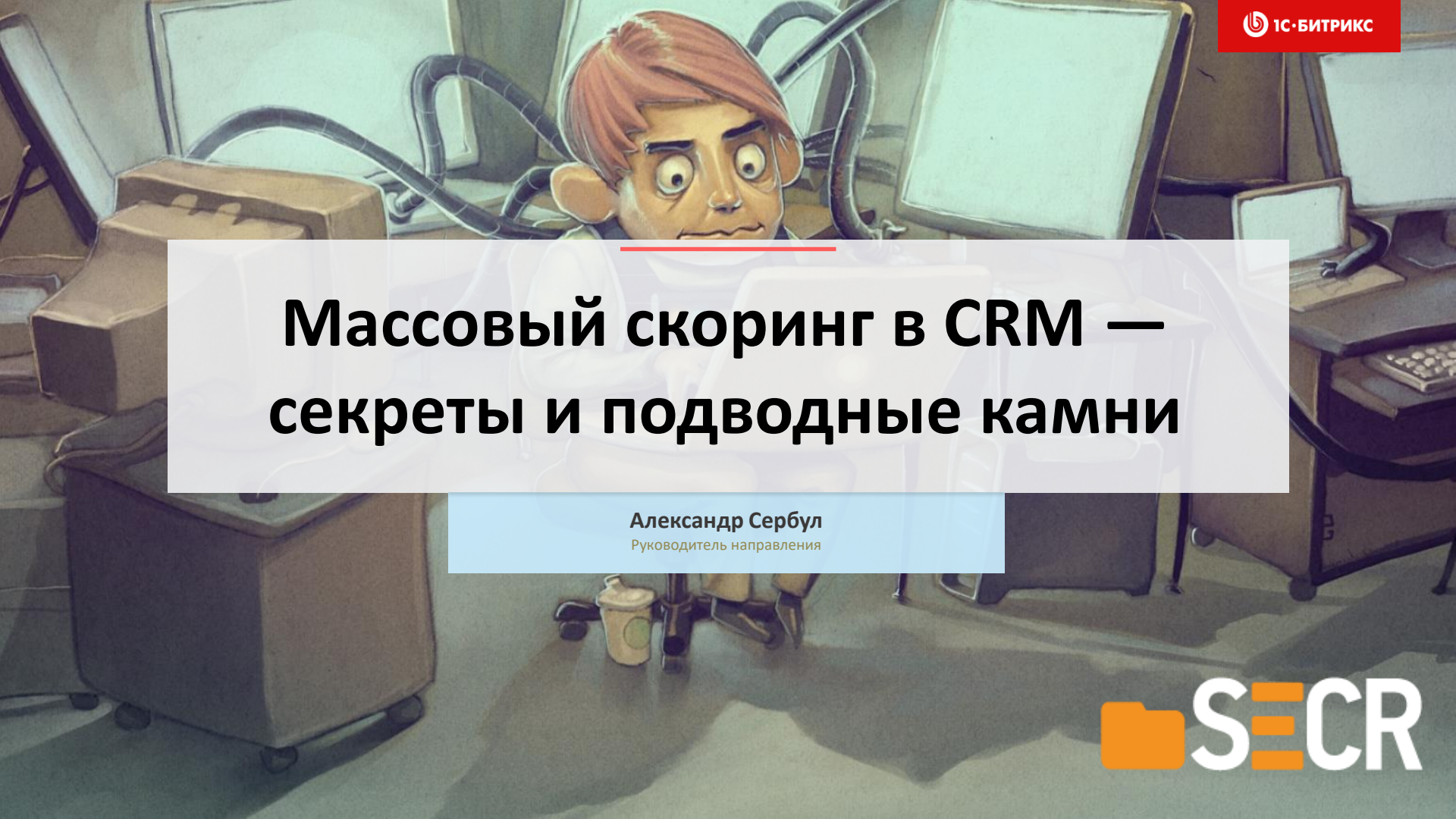


Массовый скоринг в CRM — секреты и подводные камни

Александр Сербул
Руководитель направления

О чем поговорим

1) Как мы пришли к решению сделать массовый скоринг в CRM. Потребности бизнеса, клиентов, рынка.
2) Выбор технологического стека, первый прототип.
3) Полезные новейшие возможности Amazon Web Services и, вообще, облаков, для скоринга и других применений ML.
4) Эпопея по выбору фич.
5) Несбалансированные данные – как не сойти с ума.
6) Оптимизация моделей скоринга: grid, random, байес …
7) Внедрение скоринга в продукт – интерфейс дело тонкое и интеллектуальные барьеры.
8) Статистика по боевой эксплуатации проекта.
9) О чем мечтаем.

CRM становятся все интереснее…

Лид (lead)

Некая контактная информация
Город
Канал регистрации
Число и типы писем
Число и типы звонков
Число общений в «открытых линиях»
Число и продолжительность встреч
Число задач
И многие другие

Сделка (deal)

Задача (task)

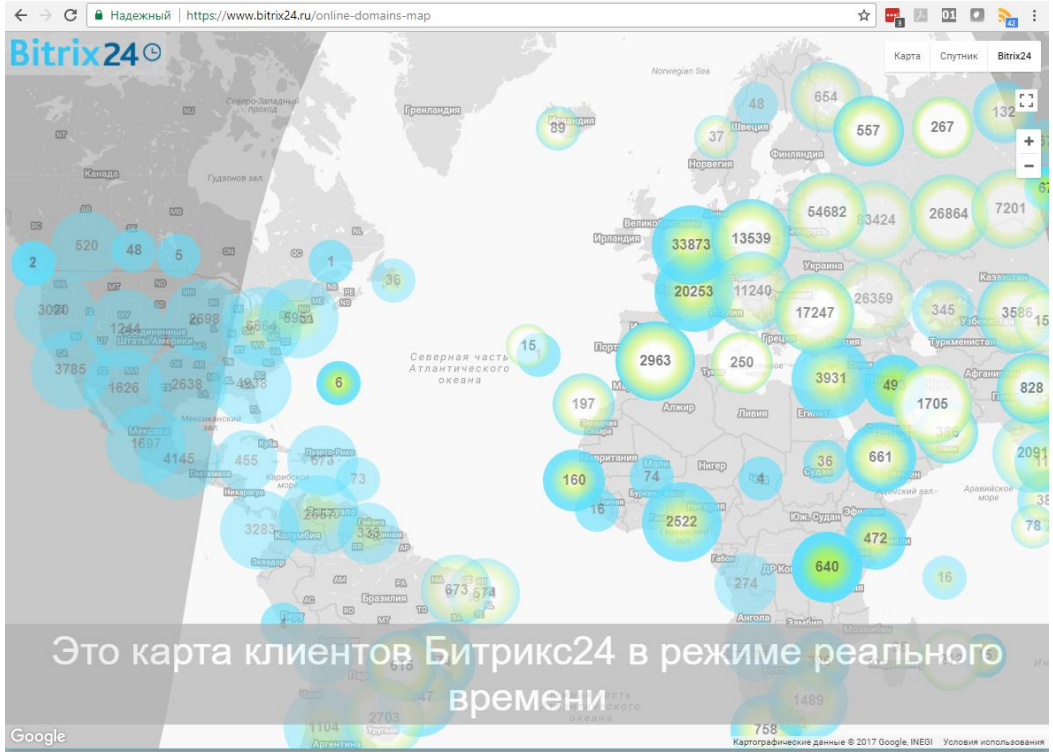
Компания (company)

Встреча (meeting)

Контакт (contact)

Каждая циферка – компания, многие - с данными в CRM

1) Миллионы компаний-клиентов с кучей данных в CRM
2) Данные не должны простаивать ☺
3) Мы поверили, что можно сделать универсальный алгоритм скоринга для разных фич, клиентов, видов бизнеса и взялись за дело …



Почему людям может быть нужен скоринг Лидов и Сделок в CRM?

- Много данных, с чего менеджеру начинать рабочий день? Геймификация.
- Аналитика – непонятно, дорого?, абстрактно. Где взять настоящих аналитиков?
- Лень думать, хочется БОЛЬШУЮ ЗЕЛЕНУЮ КНОПКУ
- Любим одну циферку больше, чем несколько. Сжатая статистика…
- Простота – залог надежности
- Лента в Facebook и другие способы принятия решений
- Верхние тарифы крупных вендоров: SalesForce, SAP …

Выбор фич – основной секрет успеха

- встречи с бизнес экспертами
- анализ хранимых в Битрикс24 CRM сущностей
- учет активностей Лида: звонки, письма, сообщения в открытую линию, встречи и т.п.
- нормализация счетчиков по интервалам
- распознавание телефонных разговоров с менеджерами, не содержащих персданных
- анализ некоторых текстов переписки по email и в открытых линиях, не содержащих персданных
- Лайки/дизлайки…

У каждого клиента – уникальный набор фич.

Number of attributes	69
Binary	6
Categorical	8
Numeric	19
Text	36

Тянем данные через REST API Битрикс24

- Пользователи порталов
- Документооборот
- Задачи
- CRM
- Авторизация

https://dev.1c-bitrix.ru/rest_help/

Как мы делали прототип …

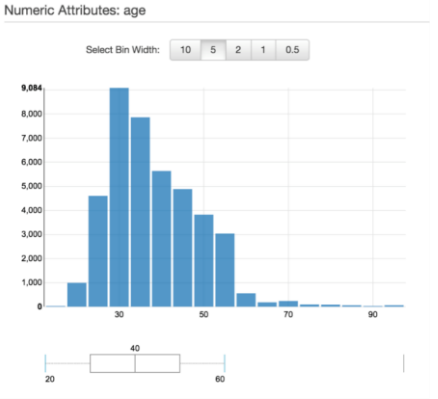
- Опыт прошлых проектов: рекомендательная система для интернет-магазинов (Apache Spark, Apache Lucene, Apache Mahout); чатботы (Deeplearning4j).
- Технологический стек компании: PHP, JavaScript, C++, Java.
- Технологический стек прототипа (август 2018): python, Jupyter Notebooks, scikit-learn, anaconda, pandas, seaborn.
- Выгрузка датасетов CRM клиентов из Битрикс24 REST API
- Первые фичи, встречи с экспертами бизнеса
- Команда: 1 человек, 2 недели.
- Трудности и методы их преодоления.

Непростой выбор предиктивной модели для скоринга

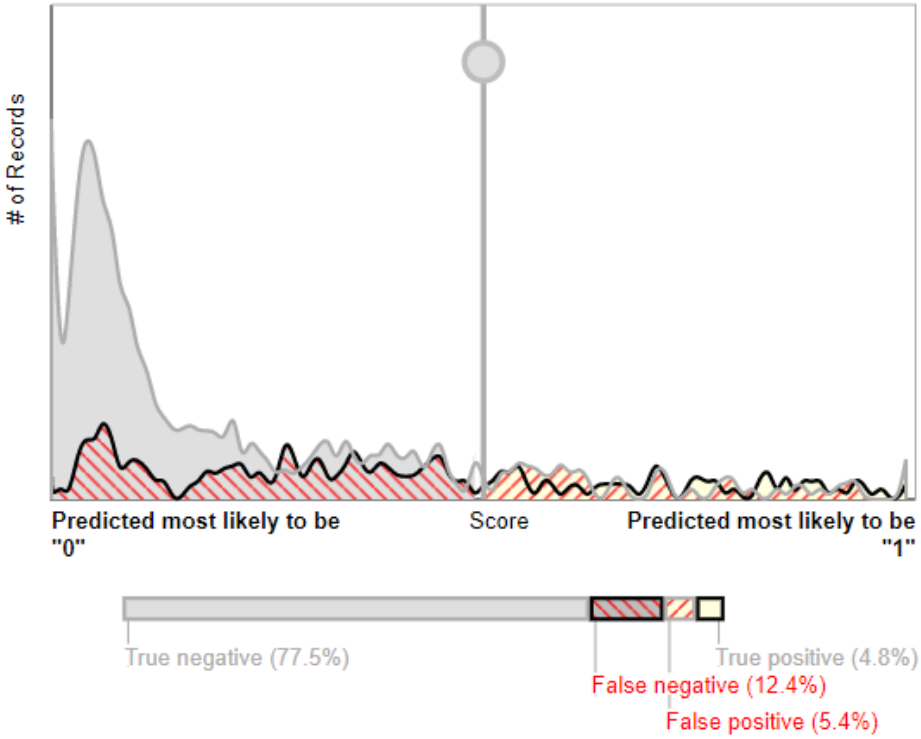
- Простые, интерпретируемые линейные модели (+ полиномиальное преобразование фич, kernel-trick)
- Машина опорных векторов
- Метод ближайших соседей
- Машина факторизации
- Деревья решений, xgboost
- Нейронная сеть

Сервис: «Amazon Machine Learning»

- простая линейная и интерпретируемая модель: логистическая регрессия
- админки для загрузки CSV, выгрузки предсказаний
- базовые статистики по фичам: min, max, средние, гистограммы
- просто отличная визуализация качества классификатора с возможностью поиграть с recall/precision
- автоматическая «подгонка» модели под размер (feature pruning)
- поддержка текстов: автосоздание корпуса, визуализация самых влияющих слов
- деплой моделей под практически любые нагрузки
- язык трансформации фич (биннинг, n-grams …)



Скоринг CRM в AWS – первые попытки



Trade-off based on score threshold 0.5

- **82% are correct**
  56 true positive
  912 true negative
- **18% are errors**
  63 false positive
  146 false negative

- 10% of the records are predicted as "1"
- 90% of the records are predicted as "0"

Save score threshold at 0.50

Advanced metrics

False positive rate **0.0646**	0
Precision **0.4706**	0
Recall **0.2772**	0
Accuracy **0.8224**	0

Несбалансированный датасет и оптимизация модели

Эксперименты на прототипе:

- “imbalanced-learn”: upsampling, downsampling
- scikit-learn: grid, random-search
- hyperopt – иногда лучше random-search, а иногда хуже ☺
- другие модели: SVM, деревья разные и пушистые, до нейронок дело не дошло пока

Что мы выбрали:

- upsampling (простой, выборкой с повторением)
- минимальный датасет: >2000 лидов или сделок
- AUC, порог вероятности 0.5, пока без настройки
- L2-регуляризация (поможет для коррелирующих фич), вопрос залу: почему не L1?


ПИТОНИСТАМ МАЛО ПЛАТЯТ...
СТАНОВИСЬ ДАТА САЕНТИСТОМ КАК Я
ТЫ ЧЁ ПЁС, Я МАТЕМАТИК
ЭТО ЗАЧЕМ?
БУДЕШЬ РАЗРАБАТЫВАТЬ ИСКУССТВЕННЫЙ ИНТЕЛЕКТ ЗА 300К/СЕК
ТАК ТЫ ЖЕ ПРОСТО РАНДОМНО ПОДБИРАЕШЬ КОЭФФИЦИЕНТЫ ПОКА КРОСС-ВАЛИДАЦИЯ НЕ ДАСТ НОРМАЛЬНЫЙ РЕЗУЛЬТАТ

Тонкости процессинга фич и деления датасета в AWS

1С·БИТРИКС

Number of files	1
Data format	CSV
Total size	3.1 MB

Data rearrangement	{
	"splitting": {
	"percentBegin": 0,
	"percentEnd": 75,
	"strategy": "random",
	"strategyParams": {
	"randomSeed": "11318_CRM_LEAD_SCORING"
	}
	}

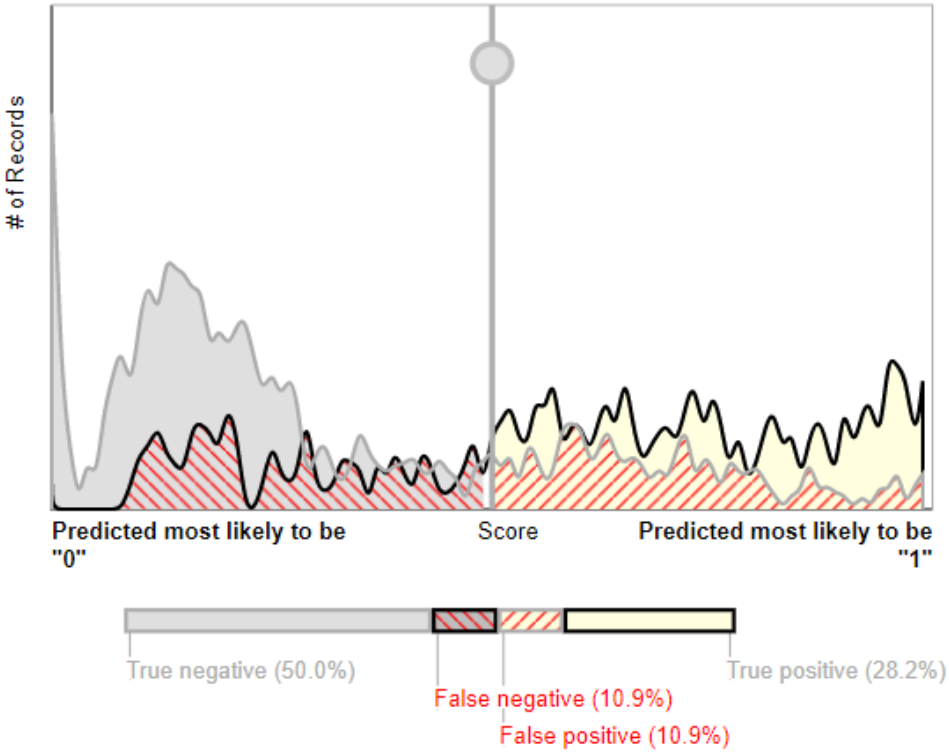
Training parameters

Maximum ML model Size	10.0 MB
Maximum number of data passes	100
Shuffle type for training data	Auto
Regularization type (amount)	L2, 1e-2 - Aggressive

Recipe

{"groups":{},"assignments":{},"outputs":["ALL_BINARY","ALL_CATEGORICAL","quantile_bin(ALL_NUMERIC,10)","lowercase(no_punct(ALL_TEXT))"]}

Скоринг CRM в Amazon Web Services – стало лучше :-)



Trade-off based on score threshold 0.5

- **78% are correct**
  457 true positive
  809 true negative
- **22% are errors**
  176 false positive
  176 false negative

- 39% of the records are predicted as "1"
- 61% of the records are predicted as "0"

Save score threshold at 0.50

Advanced metrics

False positive rate **0.1787**	0
Precision **0.722**	0
Recall **0.722**	0
Accuracy **0.7824**	0

Как поверить в модель? Я серьезно.

Эксперименты на прототипе:

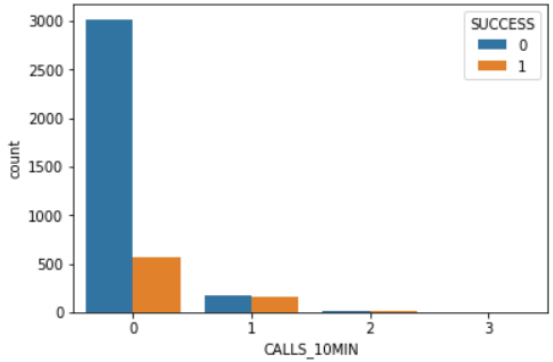
- корреляционный анализ, pearson, kendall, копулы?
- хорошенько глазками посмотреть распределения фичи и целевого признака и подумать
- деревья, качество и отсутствие адекватности прогноза

В Amazon Machine Learning:

- посмотреть силу корреляции фичи с целевым признаком

```
In [352]: #Продолжительные >10 минут звонки сильно влиют на конверсию
          sns.countplot(x='CALLS_10MIN', hue='SUCCESS', data=ds)
```

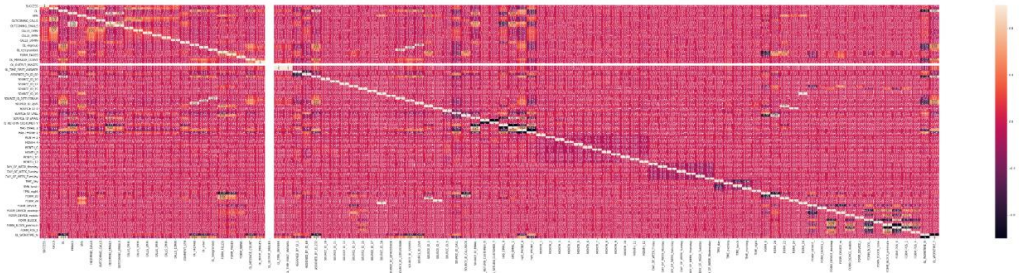
```
Out[352]: <matplotlib.axes._subplots.AxesSubplot at 0x1c71358ccc0>
```




```
In [353]: sns.countplot(x='HAS_PHONE_1', hue='SUCCESS', data=ds)
```

```
In [22]: plt.figure(figsize = (75,16))
         sns.heatmap(ds.corr(), annot=True)
```

```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x18385a3f4a8>
```



AWS: сила связи фичи с целью

1С·БИТРИКС

Numeric attributes

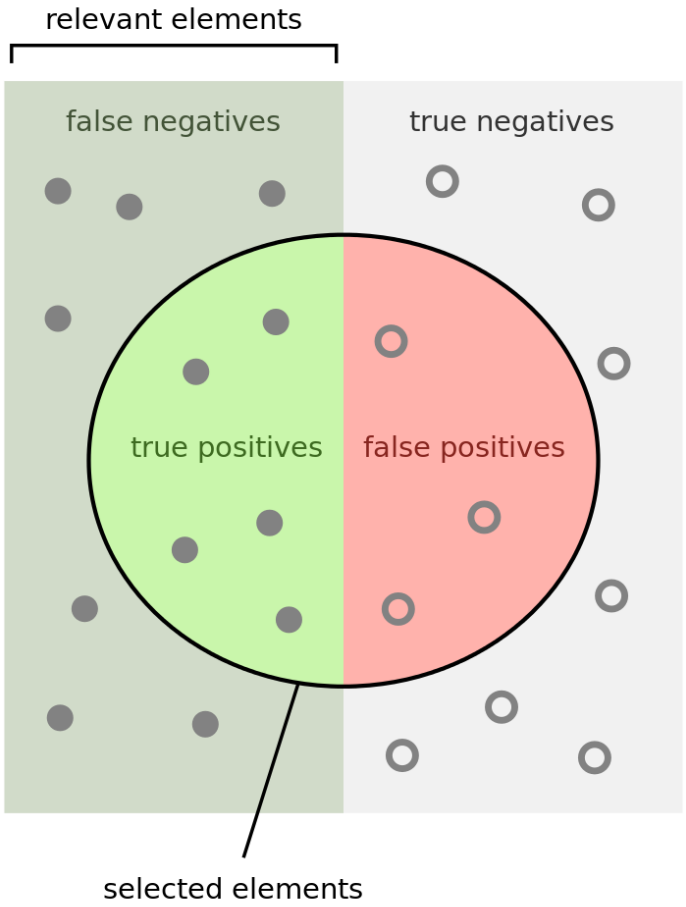
Attributes	Correlations to target *	Missing values	Invalid values	Range	Mean	Median	Preview
OUTCOMING_CALLS	0.11464	0 (0%)	2 (1%)	0 - 12	0.6726374305126621	0	
CALLS	0.06742	0 (0%)	2 (1%)	0 - 32	1.7164916615194565	1	
CALLS_10MIN	0.06301	0 (0%)	2 (1%)	0 - 3	0.13712168004941322	0	
EMAILS	0.04773	0 (0%)	2 (1%)	0 - 18	0.1340333539221742	0	
INCOMING_EMAILS	0.04427	0 (0%)	2 (1%)	0 - 9	0.09944410129709698	0	
CALLS_1MIN	0.02175	0 (0%)	2 (1%)	0 - 10	0.5762816553428042	0	
CALLS_5MIN	0.01989	0 (0%)	2 (1%)	0 - 4	0.17850525015441632	0	
OUTCOMING_EMAILS	0.01652	0 (0%)	2 (1%)	0 - 9	0.034589252625077206	0	
CALLS_0MIN	0.01506	0 (0%)	2 (1%)	0 - 17	0.3452748610253243	0	
CALLS_3MIN	0.01151	0 (0%)	2 (1%)	0 - 3	0.15441630636195183	0	

Categorical attributes

Attributes	Correlations to target	Unique values	Most frequent categories	Least frequent	Preview
HAS_PHONE	0.06742	3	1	HAS_PHONE	
SOURCE_ID	0.04247	17	CALL	17	
HAS_EMAIL	0.0303	3	0	HAS_EMAIL	
FORM	0.01346	5	0	24	
OL_WORKTIME	0.00633	3	N	OL_WORKTIME	
FORM_DEVICE	0.00513	5		mobile	
TIME	0.00438	6	day	TIME	
ASSIGNED_BY_ID	0.00304	4	1	ASSIGNED_BY_ID	
MONTH	0.0019	13	3	MONTH	
FORM_POS	0.00172	3	0	1	

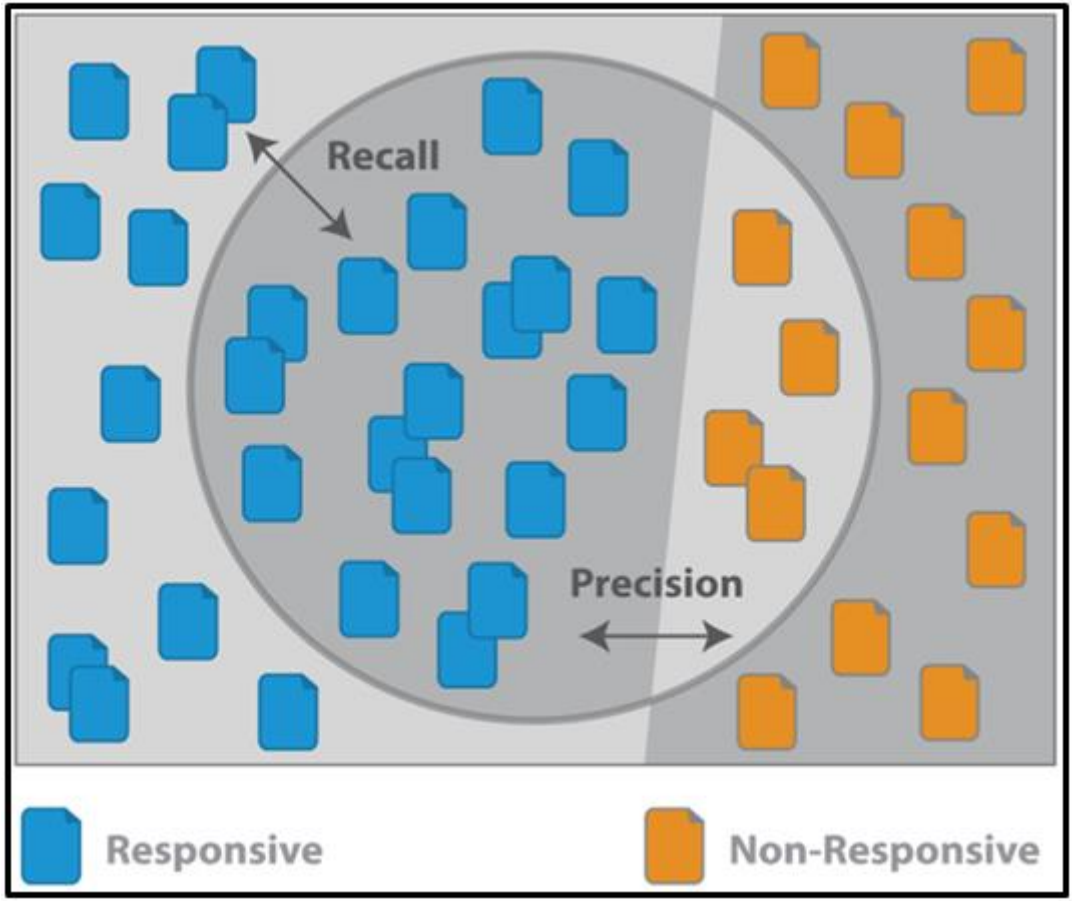
Как контролировать качество в «скоринге»?

- Понять физику и химию Precision и Recall
- Не увлекаться циферками: ROC AUC и т.п. Балансировка и AUC.
- Очень неудачные термины: ошибка 1-2 рода, основная и альтернативные гипотезы
- Посмотреть на полученный прогноз глазами менеджера, склонного к насилию, знающего ваш домашний адрес



Метрики – только самые нужные

- **Precision** – процент «мусора» в ответе
  - Точность работы поисковика
- **Recall** – сколько данных удалось достать в ответе?
  - Сколько из того, что был должен дать реально дал



Метрики – только самые нужные

Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

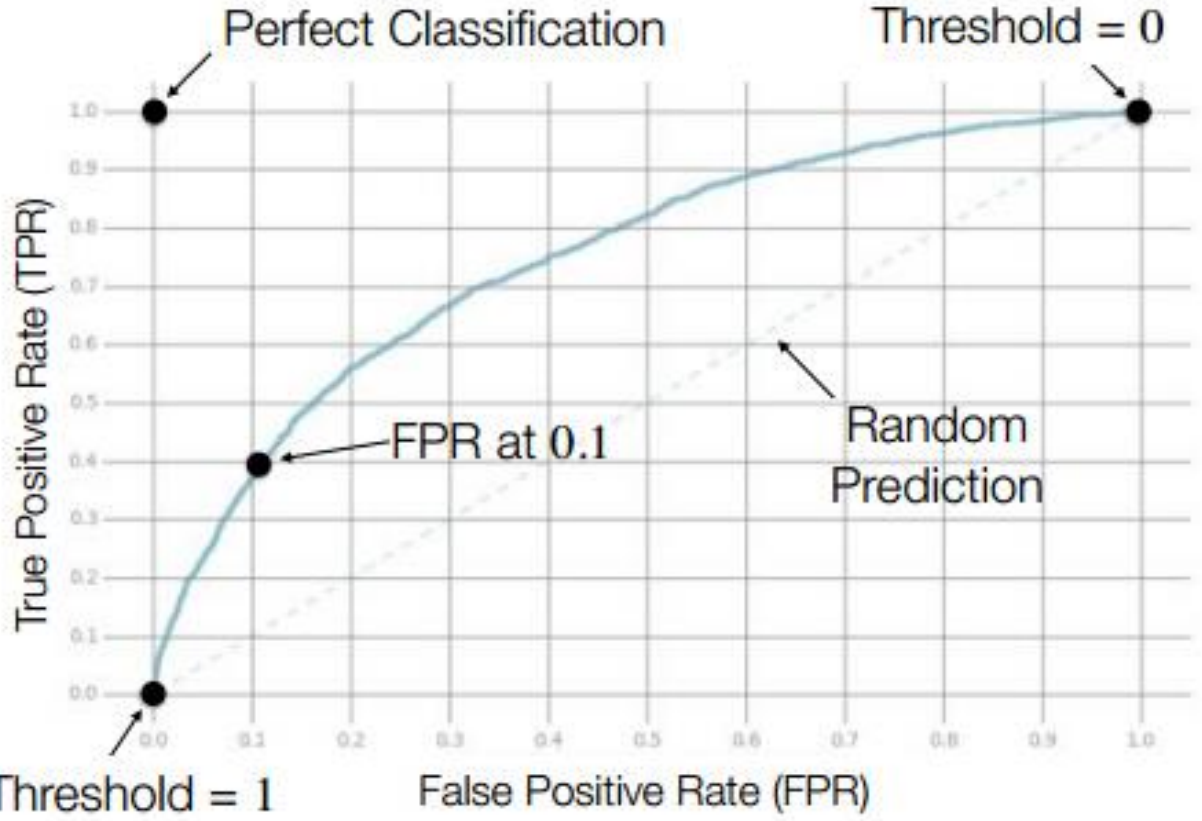
TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

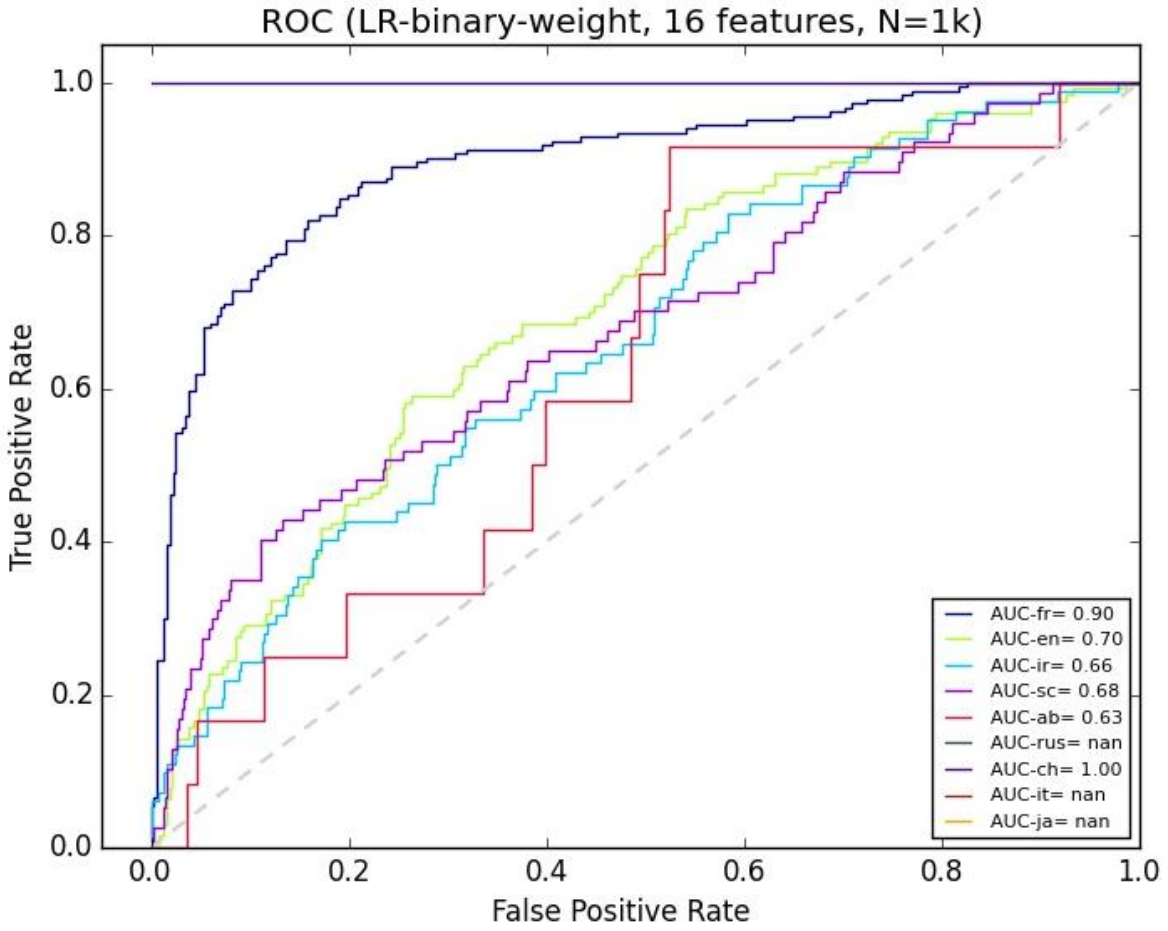
Метрики – только самые нужные

- AUC (area under curve)



Метрики – только самые нужные

- Неадекватен на несбалансированных классах

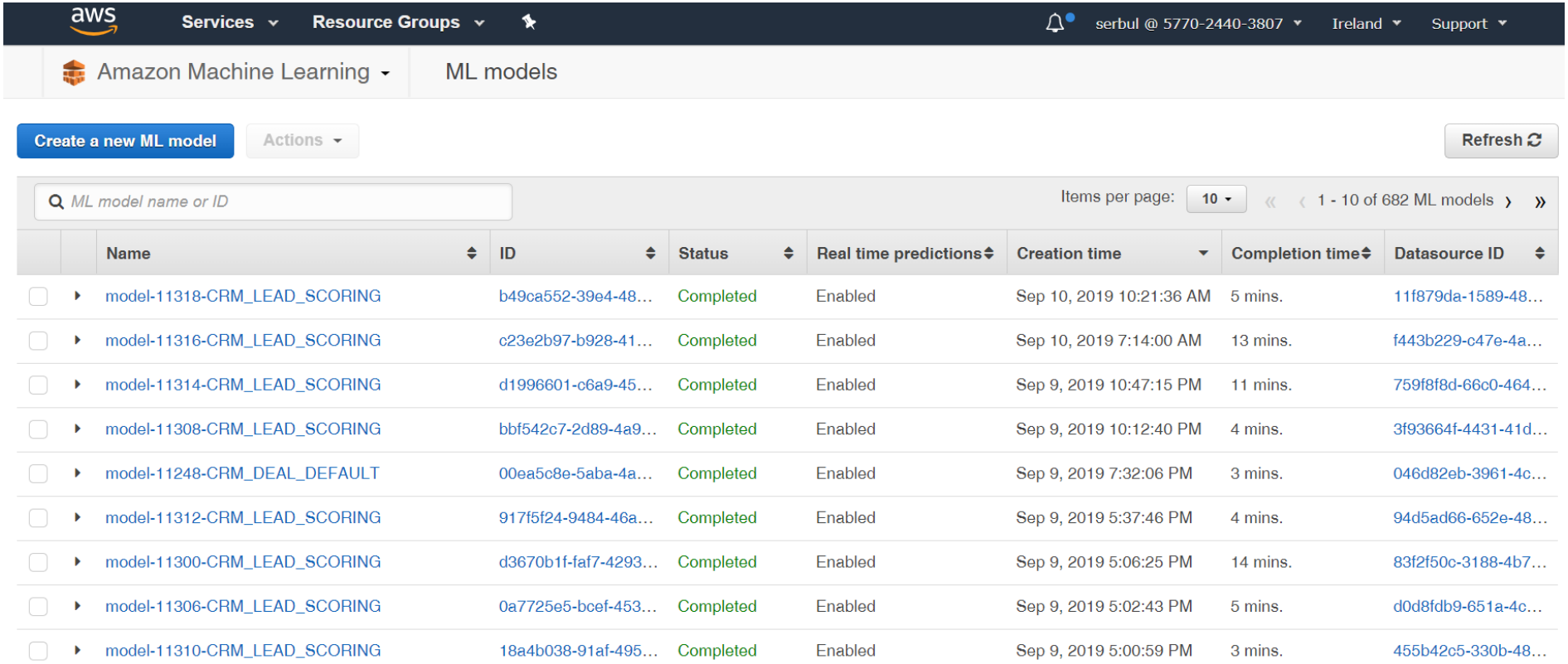


Сравнение со случайным классификатором

- Адаптировать метрики под сильно несбалансированный датасет
- Почитать recall, precision, f1 и сравнить с полученным моделью
- Искусство, никакой науки ☺

Статистика и факты «с боя» - нагрузка

Почти 700 моделей буквально в считанные дни поле запуска на ограниченную аудиторию … 1 модель – 1 компания



Name	ID	Status	Real time predictions	Creation time	Completion time	Datasource ID
model-11318-CRM_LEAD_SCORING	b49ca552-39e4-48…	Completed	Enabled	Sep 10, 2019 10:21:36 AM	5 mins.	11f879da-1589-48…
model-11316-CRM_LEAD_SCORING	c23e2b97-b928-41…	Completed	Enabled	Sep 10, 2019 7:14:00 AM	13 mins.	f443b229-c47e-4a…
model-11314-CRM_LEAD_SCORING	d1996601-c6a9-45…	Completed	Enabled	Sep 9, 2019 10:47:15 PM	11 mins.	759f8f8d-66c0-464…
model-11308-CRM_LEAD_SCORING	bbf542c7-2d89-4a9…	Completed	Enabled	Sep 9, 2019 10:12:40 PM	4 mins.	3f93664f-4431-41d…
model-11248-CRM_DEAL_DEFAULT	00ea5c8e-5aba-4a…	Completed	Enabled	Sep 9, 2019 7:32:06 PM	3 mins.	046d82eb-3961-4c…
model-11312-CRM_LEAD_SCORING	917f5f24-9484-46a…	Completed	Enabled	Sep 9, 2019 5:37:46 PM	4 mins.	94d5ad66-652e-48…
model-11300-CRM_LEAD_SCORING	d3670b1f-faf7-4293…	Completed	Enabled	Sep 9, 2019 5:06:25 PM	14 mins.	83f2f50c-3188-4b7…
model-11306-CRM_LEAD_SCORING	0a7725e5-bcef-453…	Completed	Enabled	Sep 9, 2019 5:02:43 PM	5 mins.	d0d8fdb9-651a-4c…
model-11310-CRM_LEAD_SCORING	18a4b038-91af-495…	Completed	Enabled	Sep 9, 2019 5:00:59 PM	3 mins.	455b42c5-330b-48…

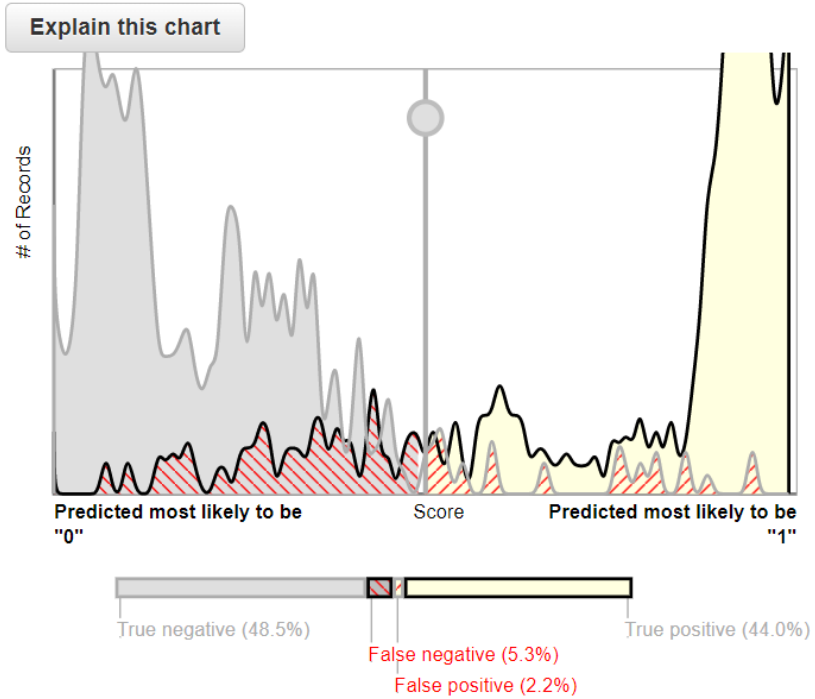
Статистика и факты «с боя» - детали обучения модели

Средний размер моделей: 3.7 МБ

```
Message-ID: <Tue Sep 10 07:32:09 UTC 2019_577024403807-pr-b49ca552-39e4-4899-9e26-48bc0fccb522/userlog/577024403807-pr-b49ca552-39e4-4899-9e26-48bc0f
MIME-Version: 1.0
Content-Type: multipart/mixed;
        boundary="----=_Part_1727_1967990270.1568100729560"

------=_Part_1727_1967990270.1568100729560
Amazon Machine Learning


------=_Part_1727_1967990270.1568100729560
19/09/10 07:27:58 INFO: Begin training.
19/09/10 07:28:07 INFO: initial-training: l1=0.0 l2=0.01 likelihood-function=logreg max-passes=100 max-model-size=10485760 (10.00 MB) readable=false
19/09/10 07:28:08 INFO: learner-id=1050 model-configuration:       learning-rate=0.01
19/09/10 07:28:08 INFO: learner-id=1050 model-convergence:         negative-log-likelihood=1.000000e+00 (delta=1.000000e+00) is-converged=no
19/09/10 07:28:08 INFO: learner-id=1050 active-features:           updates=0000000000 min=0000000 max=0000000 mean=0000000 total-sum=0000000000
19/09/10 07:28:08 INFO: learner-id=1050 active-features-quantiles: quantile-10=0000000 quantile-50=0000000 quantile-90=0000000
19/09/10 07:28:08 INFO: learner-id=1050 model-status:              model-size=0 (0.00 MB) #params=0 #pruning-calls=0000000000
```

```
finished-training: total-passes=100 valid-records=414700 invalid-records=0
best learner:
learner-id=3535 model-configuration:      learning-rate=1.0
learner-id=3535 model-performance:        accuracy=0.9276 recall=0.8858 precision=0.9673 f1-score=0.9248 auc=0.9772
learner-id=3535 model-convergence:        negative-log-likelihood=2.282126e-01 (delta=1.000000e+00) is-converged=no
learner-id=3535 active-features:          updates=0000414700 min=0000068 max=0001790 mean=0000118 total-sum=0048564600
learner-id=3535 active-features-quantiles: quantile-10=0000076 quantile-50=0000094 quantile-90=0000110
learner-id=3535 model-status:             model-size=1221024 (1.16 MB) #params=12719 #pruning-calls=0000000004


final consolidation:

saved final learner:
learner-id=3535 model-configuration:      learning-rate=1.0
learner-id=3535 model-convergence:        negative-log-likelihood=2.282126e-01 (delta=1.000000e+00) is-converged=no
learner-id=3535 active-features:          updates=0000414700 min=0000068 max=0001790 mean=0000118 total-sum=0048564600
learner-id=3535 active-features-quantiles: quantile-10=0000076 quantile-50=0000094 quantile-90=0000110
learner-id=3535 model-status:             model-size=1217184 (1.16 MB) #params=12679 #pruning-calls=0000000005

initial-training: finished
Foorprint size is 1698076
Training complete.
```

Качество моделей, в основном, «подозрительно» хорошее



Disable real time predictions to update the threshold.

Trade-off based on score threshold 0.5 Reset score threshold (0.5)

- **92% are correct**
  650 true positive
  716 true negative
- **8% are errors**
  33 false positive
  78 false negative

- 46% of the records are predicted as "1"
- 54% of the records are predicted as "0"

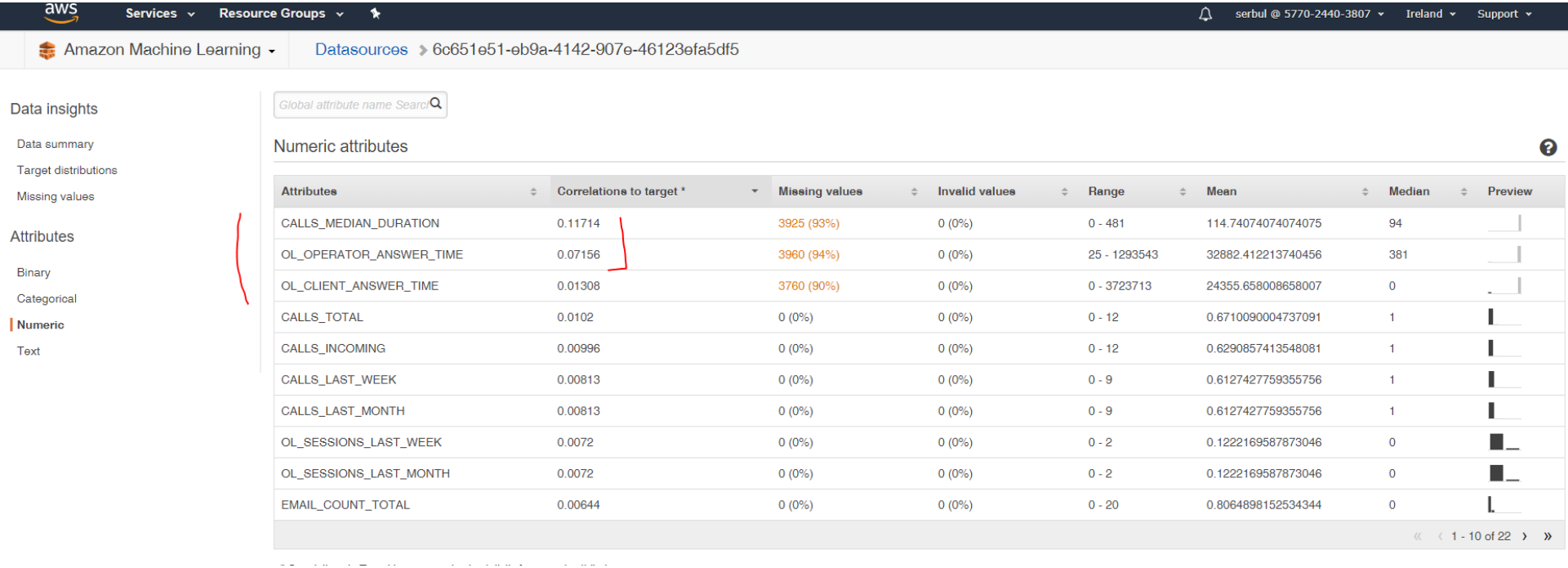
Save score threshold at 0.50

Advanced metrics

False positive rate **0.0441**	0		1
Precision **0.9517**	0		1
Recall **0.8929**	0		1
Accuracy **0.9248**	0		1

Интересные корреляции

Коммуникации - звонки



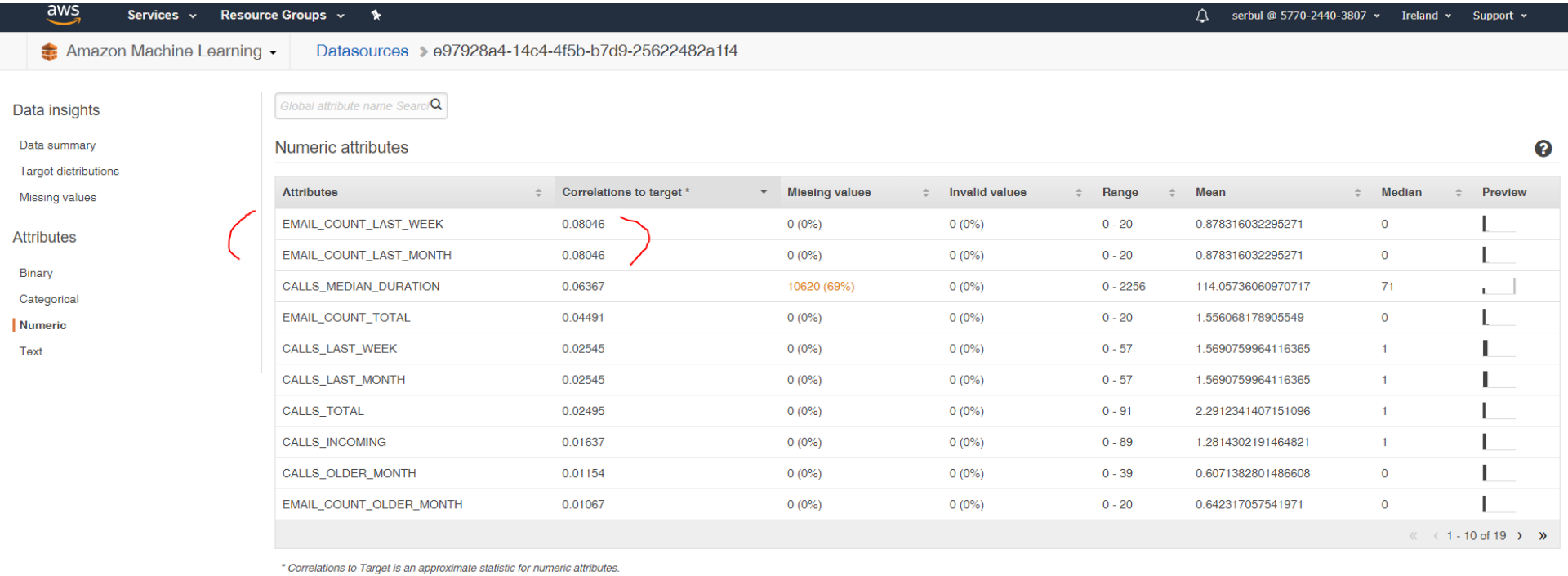
Attributes	Correlations to target *	Missing values	Invalid values	Range	Mean	Median	Preview
CALLS_MEDIAN_DURATION	0.11714	3925 (93%)	0 (0%)	0 - 481	114.74074074074075	94	
OL_OPERATOR_ANSWER_TIME	0.07156	3960 (94%)	0 (0%)	25 - 1293543	32882.412213740456	381	
OL_CLIENT_ANSWER_TIME	0.01308	3760 (90%)	0 (0%)	0 - 3723713	24355.658008658007	0	
CALLS_TOTAL	0.0102	0 (0%)	0 (0%)	0 - 12	0.6710090004737091	1	
CALLS_INCOMING	0.00996	0 (0%)	0 (0%)	0 - 12	0.6290857413548081	1	
CALLS_LAST_WEEK	0.00813	0 (0%)	0 (0%)	0 - 9	0.6127427759355756	1	
CALLS_LAST_MONTH	0.00813	0 (0%)	0 (0%)	0 - 9	0.6127427759355756	1	
OL_SESSIONS_LAST_WEEK	0.0072	0 (0%)	0 (0%)	0 - 2	0.1222169587873046	0	
OL_SESSIONS_LAST_MONTH	0.0072	0 (0%)	0 (0%)	0 - 2	0.1222169587873046	0	
EMAIL_COUNT_TOTAL	0.00644	0 (0%)	0 (0%)	0 - 20	0.8064898152534344	0	

1 - 10 of 22

Интересные корреляции

1С·БИТРИКС

Коммуникации - почта



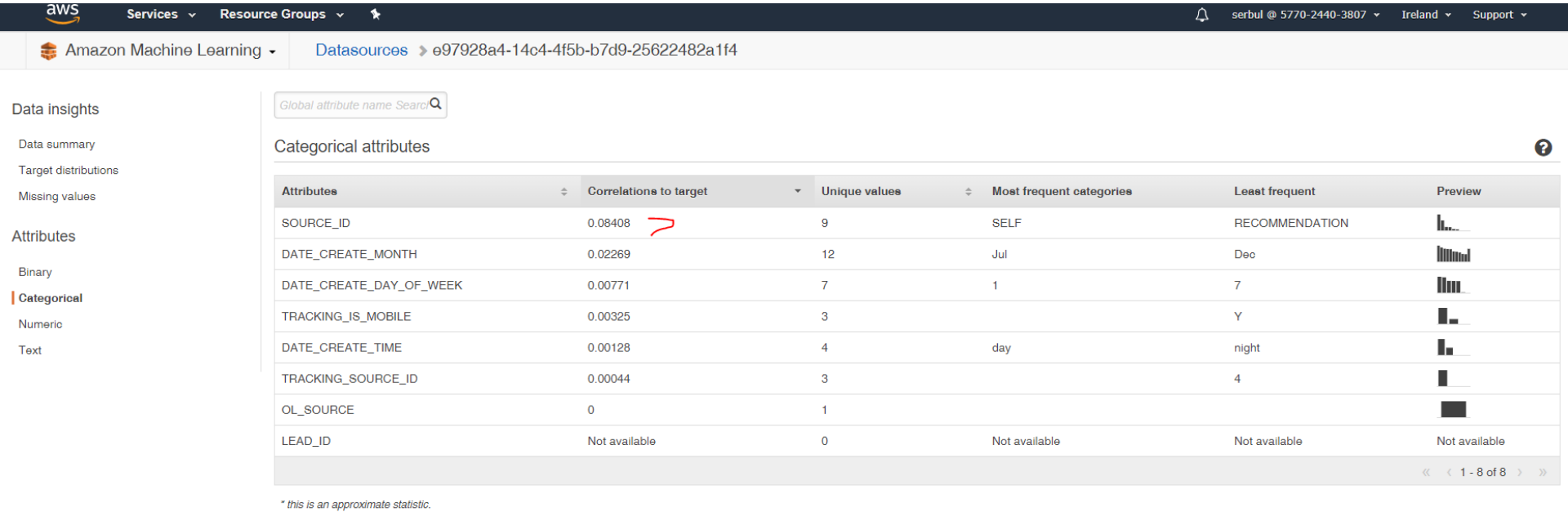
Attributes	Correlations to target *	Missing values	Invalid values	Range	Mean	Median	Preview
EMAIL_COUNT_LAST_WEEK	0.08046	0 (0%)	0 (0%)	0 - 20	0.878316032295271	0	
EMAIL_COUNT_LAST_MONTH	0.08046	0 (0%)	0 (0%)	0 - 20	0.878316032295271	0	
CALLS_MEDIAN_DURATION	0.06367	10620 (69%)	0 (0%)	0 - 2256	114.05736060970717	71	
EMAIL_COUNT_TOTAL	0.04491	0 (0%)	0 (0%)	0 - 20	1.556068178905549	0	
CALLS_LAST_WEEK	0.02545	0 (0%)	0 (0%)	0 - 57	1.5690759964116365	1	
CALLS_LAST_MONTH	0.02545	0 (0%)	0 (0%)	0 - 57	1.5690759964116365	1	
CALLS_TOTAL	0.02495	0 (0%)	0 (0%)	0 - 91	2.2912341407151096	1	
CALLS_INCOMING	0.01637	0 (0%)	0 (0%)	0 - 89	1.2814302191464821	1	
CALLS_OLDER_MONTH	0.01154	0 (0%)	0 (0%)	0 - 39	0.6071382801486608	0	
EMAIL_COUNT_OLDER_MONTH	0.01067	0 (0%)	0 (0%)	0 - 20	0.642317057541971	0	

1 - 10 of 19

** Correlations to Target is an approximate statistic for numeric attributes.*

Интересные корреляции

Источник лида



Интересные корреляции

Интересный эксперимент с текстами без персданных

Amazon Machine Learning ▾ Datasources › 14dcf2a5-49c4-4fe6-9be6-6c9d9c2ff2a3

Data insights

- Data summary
- Target distributions
- Missing values

Attributes

- Binary
- Categorical
- Numeric
- Text

Global attribute name Search

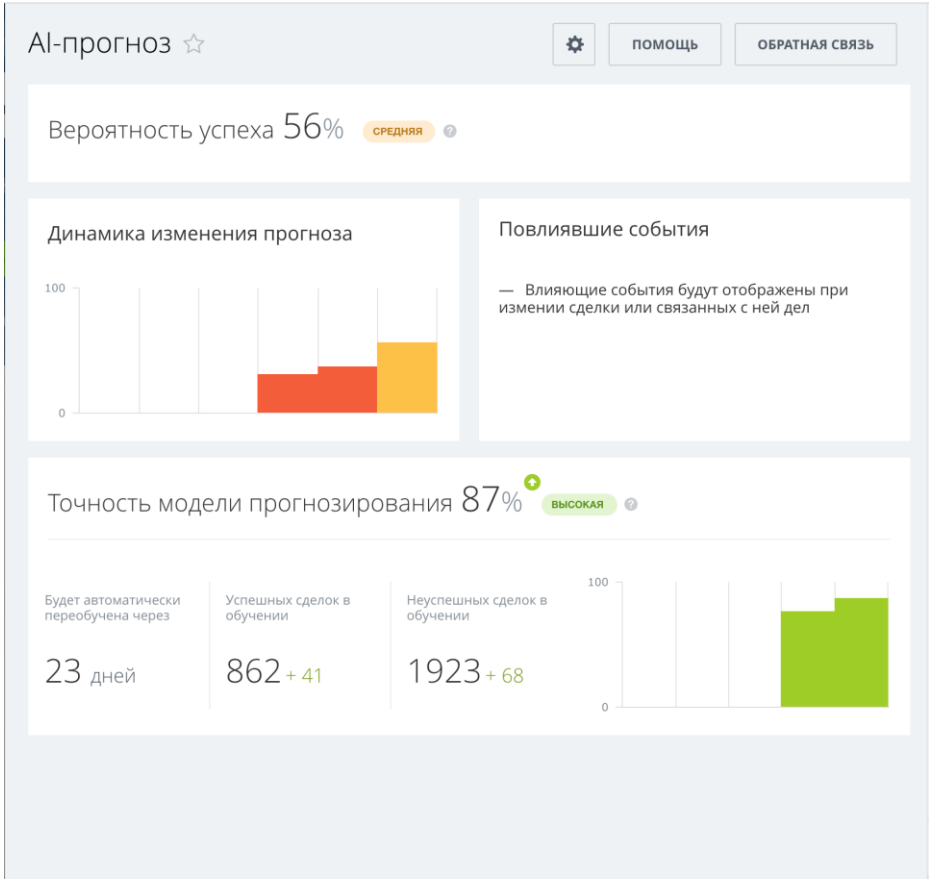
Text attributes

Attributes	Correlations to target *	Total words	Unique words	Words in attribute (range)	Word length (range)	Most prominent words
EMAIL_OPERATOR_TEXT	0.21996	136911	3297	0 - 1000	1 - 40	заказ, отправк ...
COMMENTS	0.06389	51484	2847	0 - 68	2 - 16	клиент, связ ...
UF_CRM_1477557412	0.04154	3284	603	0 - 30	2 - 15	заказ, клиент ...
TITLE	0.02939	26957	10700	0 - 7	2 - 14	заказ, центр ...
OL_CLIENT_MESSAGE_TEXT	0.01769	9299	1806	0 - 588	2 - 33	заполн, форм ...
OL_OPERATOR_MESSAGE_TEXT	0.01683	7735	995	0 - 116	2 - 16	добр, ден ...
EMAIL_CLIENT_TEXT	0.00879	28703	2399	0 - 3648	1 - 704	номер, оплач ...
UF_CRM_5A8BCDE8195D5	0.00045	17	15	0 - 7	3 - 9	заказ, сдела ...
UF_CRM_DEAL_UID	0	0	Not available	0 - 0	0 - 0	Not available
UF_CRM_DEAL_STATE	0	0	Not available	0 - 0	0 - 0	Not available

« ‹ 1 - 10 of 15 › »

* Correlations to Target is an approximate statistic for text attributes.

Сервис скоринга – глазами клиента



Дальнейшие шаги

- Оптимизация моделей внутри Amazon Sage Maker – random и байесовский поиск на кластерах.
- Активное использование ClickHouse и python внутри компании. Внутренние скоринги для бизнеса и маркетинга.
- Возможно частичный переход на более мощные модели внутри Amazon Sage Maker.
- Развитие продуктовой аналитики и предиктивного маркетинга на основе данных (которых все больше и больше и больше и больше).

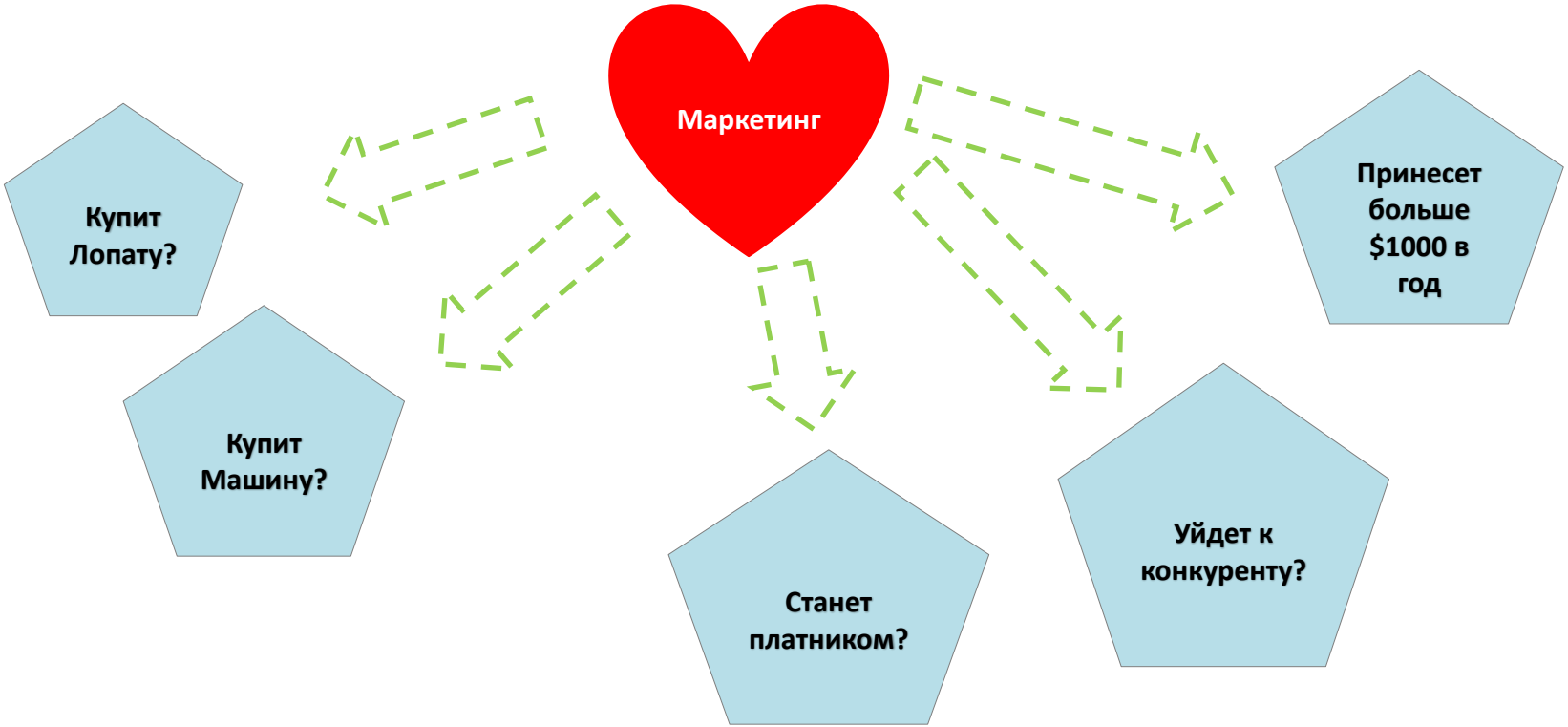
Сервис «Amazon Sage Maker»

- Немало встроенных МАСШТАБИРУЕМЫХ алгоритмов
- Поддержка работы с Jupiter Notebooks (kernels: Python 2 and 3, Apache MXNet, TensorFlow, and PySpark)
- Авто-масштабирование, развертывание, A/B-тестирование
- Оплата только за хостинг железа для моделей
- Можно поднимать машины с GPU

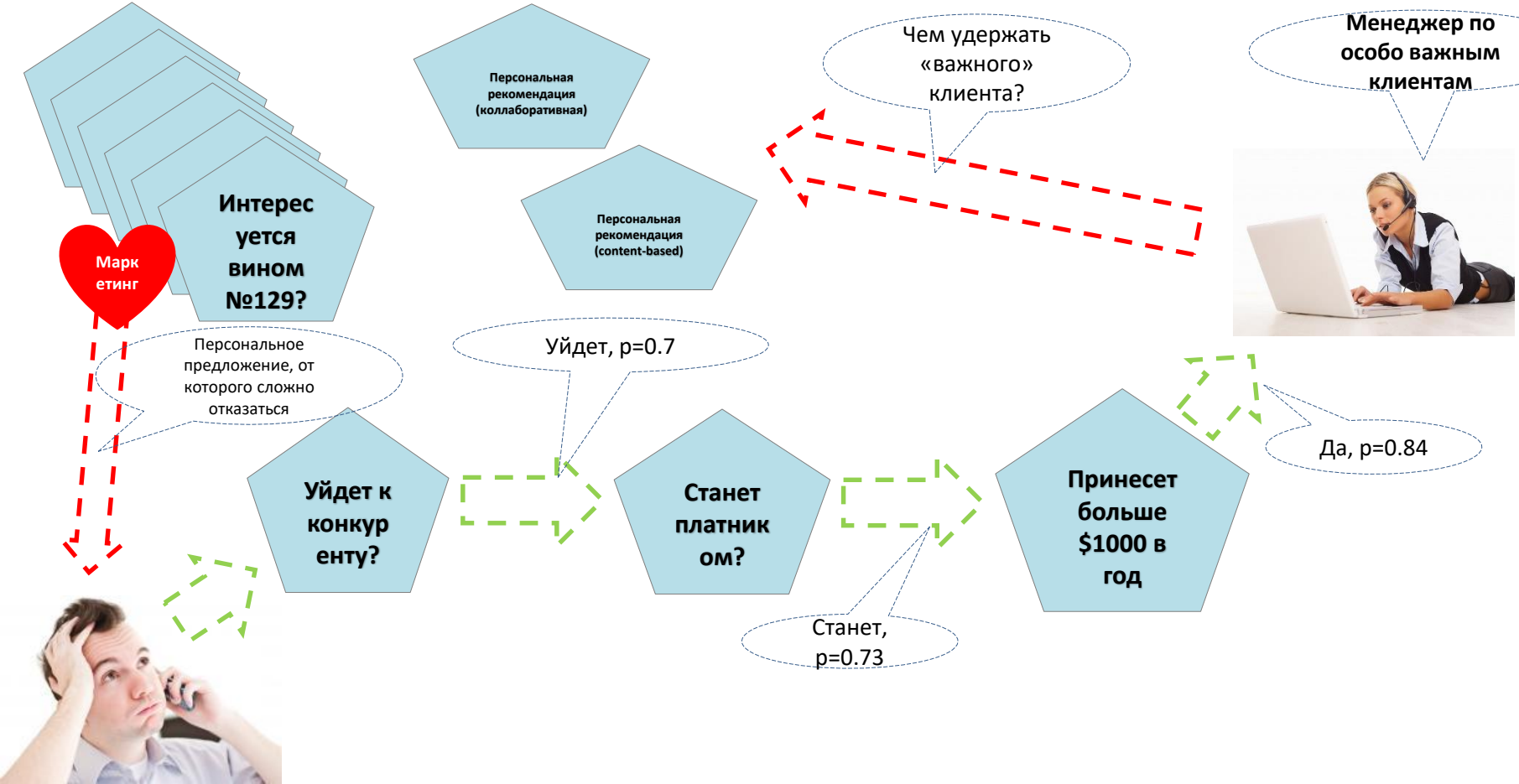
Using Built-in Algorithms

- Common Information
- Linear Learner
- Factorization Machines
- XGBoost Algorithm
- Image Classification Algorithm
- Object Detection Algorithm
- Sequence to Sequence (seq2seq)
- K-Means Algorithm
- Principal Component Analysis (PCA)
- Latent Dirichlet Allocation (LDA)
- Neural Topic Model (NTM)
- DeepAR Forecasting
- BlazingText
- Random Cut Forest
- K-Nearest Neighbors

Больше моделей, хороших и понятных!



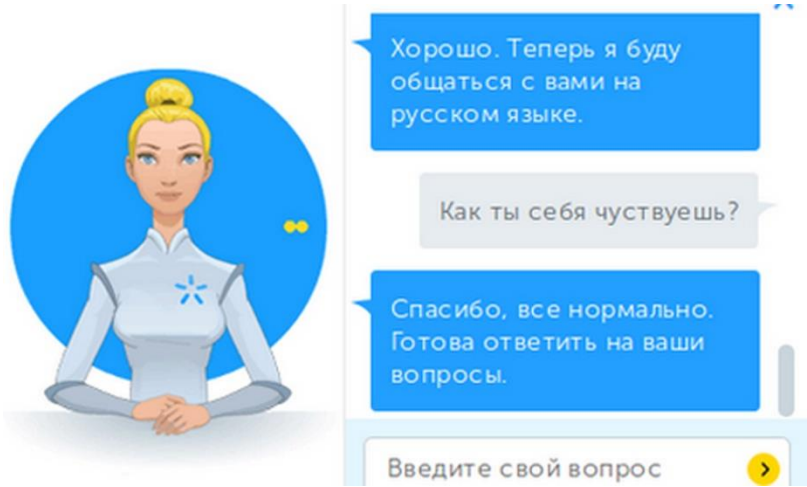
Внедрение ML в продукты для оптимизации продаж



Итоги

- Простая модель (логистическая регрессия) – работает, на удивление, хорошо
- Правильные интерпретируемые фичи – ключ к успеху и универсальным моделям
- Перед предиктивными моделями нужна хорошая, понятная аналитика и интенсивные коммуникации с экспертами бизнеса
- Сейчас довольно просто развернуть подобный массовый ML-сервис в облаке, например Amazon, Google в разумное время

Чатботы

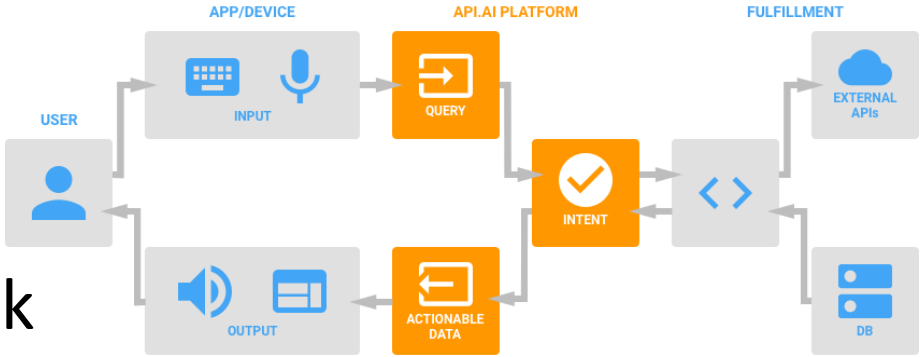

Хорошо. Теперь я буду общаться с вами на русском языке.
Как ты себя чуствуешь?
Спасибо, все нормально. Готова ответить на ваши вопросы.
Введите свой вопрос

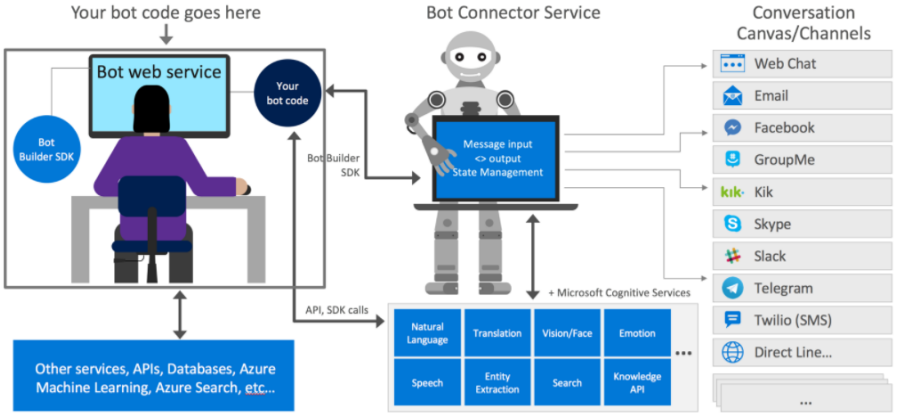
Чатботы – технологии «попроще»

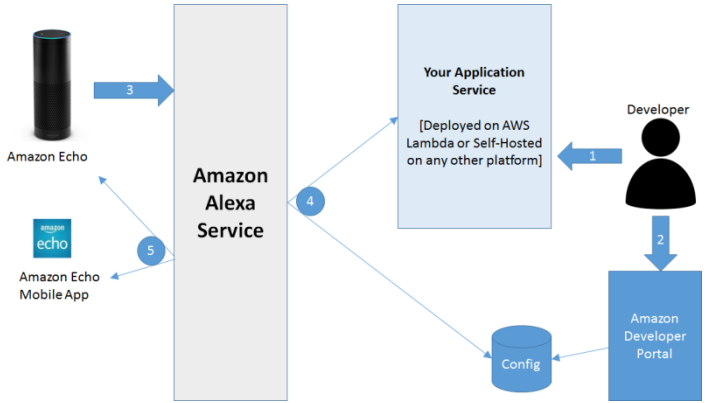
- Регулярные выражения
- Примитивные «движки» с правилами
- Заполнение цепочки форм
- Вычленение фактов из теста и выполнение действий
- api.ai – простой движок на веб-хуках

Чатботы – фреймворки

- api.ai
- Amazon Alexa
- Microsoft Bot Framework



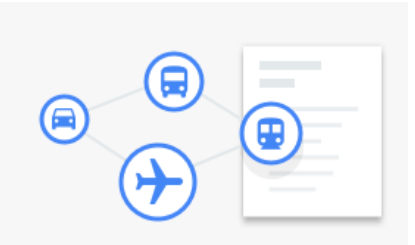




Чатботы – фреймворки, элементы

Google Cloud Natural Language API, Amazon Comprehend

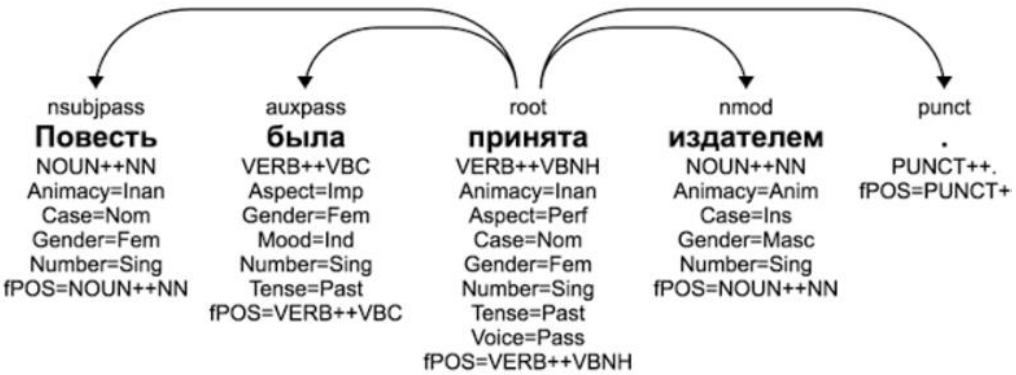
- Анализ тональности текста
- Выявление намерения (категория)
- Вычленение сущностей (личности, места, даты и т.п.)
- Синтаксический разбор (части речи)
- Ключевые слова
- Определение языков



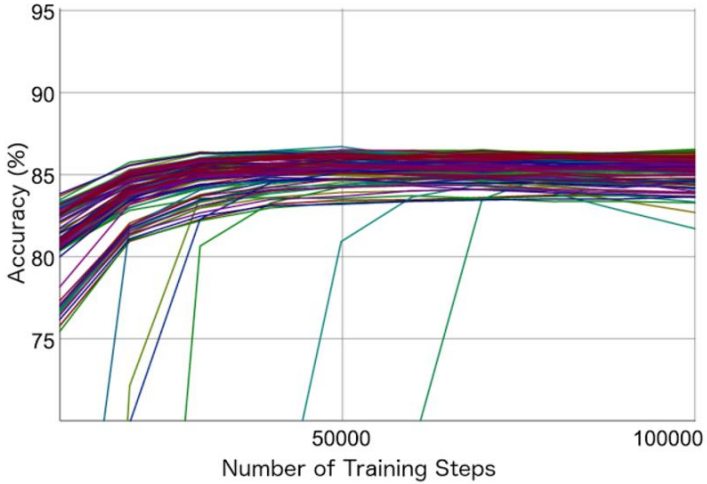
Чатботы – элементы

Google SyntaxNet, Parsey’s Cousins

>40 языков



Parse trees showing dependency labels, parts of speech, and morphology.

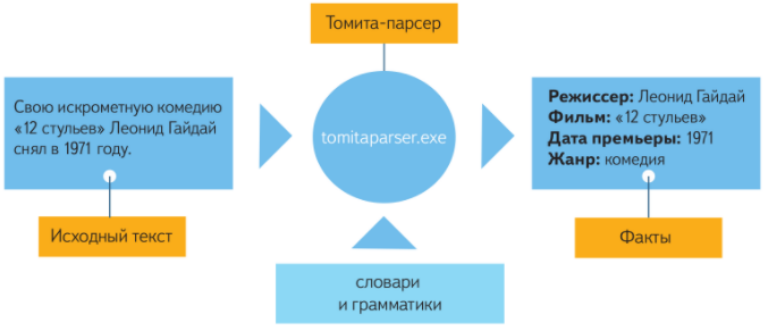


Чатботы – элементы

Яндекс, Томита.Парсер. Вычленение фактов.

текст	*Иван*	*родился*	*в*	*Нижнем*	*Новгороде*
леммы	иван	родиться	в	нижний	новгород
грамматические признаки	S, persn, nom, sg, m, anim	V, praet, sg, indic, m, ipf, intr	PR	A, abl, sg, plen, m, n	S, geo, abl, sg, m, inan
kw-type		born		city	
терминалы	AnyWord<gram="persn">	Verb<kwtype="born">	"в"	Noun<kwtype="city">	
нетерминалы	Person	Born	"в"	City	
интерпретация (поля факта)	BornFact.Person			BornFact.Place	

текст	*Михаил*	*появился*	*на*	*свет*	*в*	*Петербурге*
леммы	Михаил	появиться	на	свет	в	Санкт-Петербург
грамматические признкаи	S, persn, nom, sg, m, anim	V, praet, sg, indic, m, pf, intr	PR	S, nom, acc, sg, m, inan	PR	S, geo, abl, sg, m, inan
kw-type		born				city
терминалы	AnyWord<gram="persn">	Verb<kwtype="born">			"в"	Noun<kwtype="city">
нетерминалы	Person	Born			"в"	City
интерпретация (поля факта)	BornFact.Person					BornFact.Place



Deeppavlov.ai

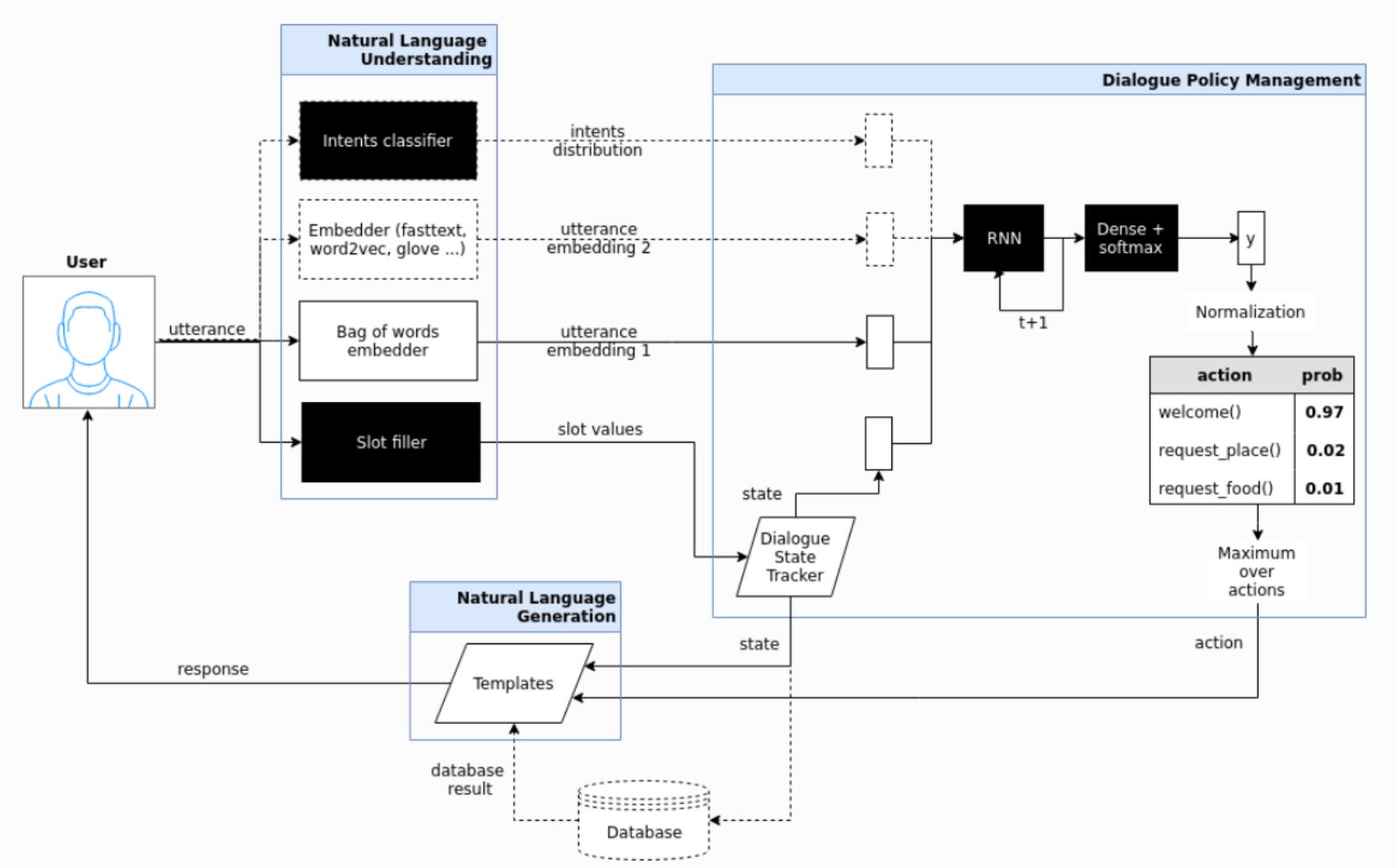
- https://deeppavlov.ai
- http://ipavlov.ai
- Очень интересные, мощные архитектуры моделей и нейросетей
- Есть готовые «русскоязычные» модели
- Отличная поддержка на форуме: https://forum.ipavlov.ai

iPavlov.ai

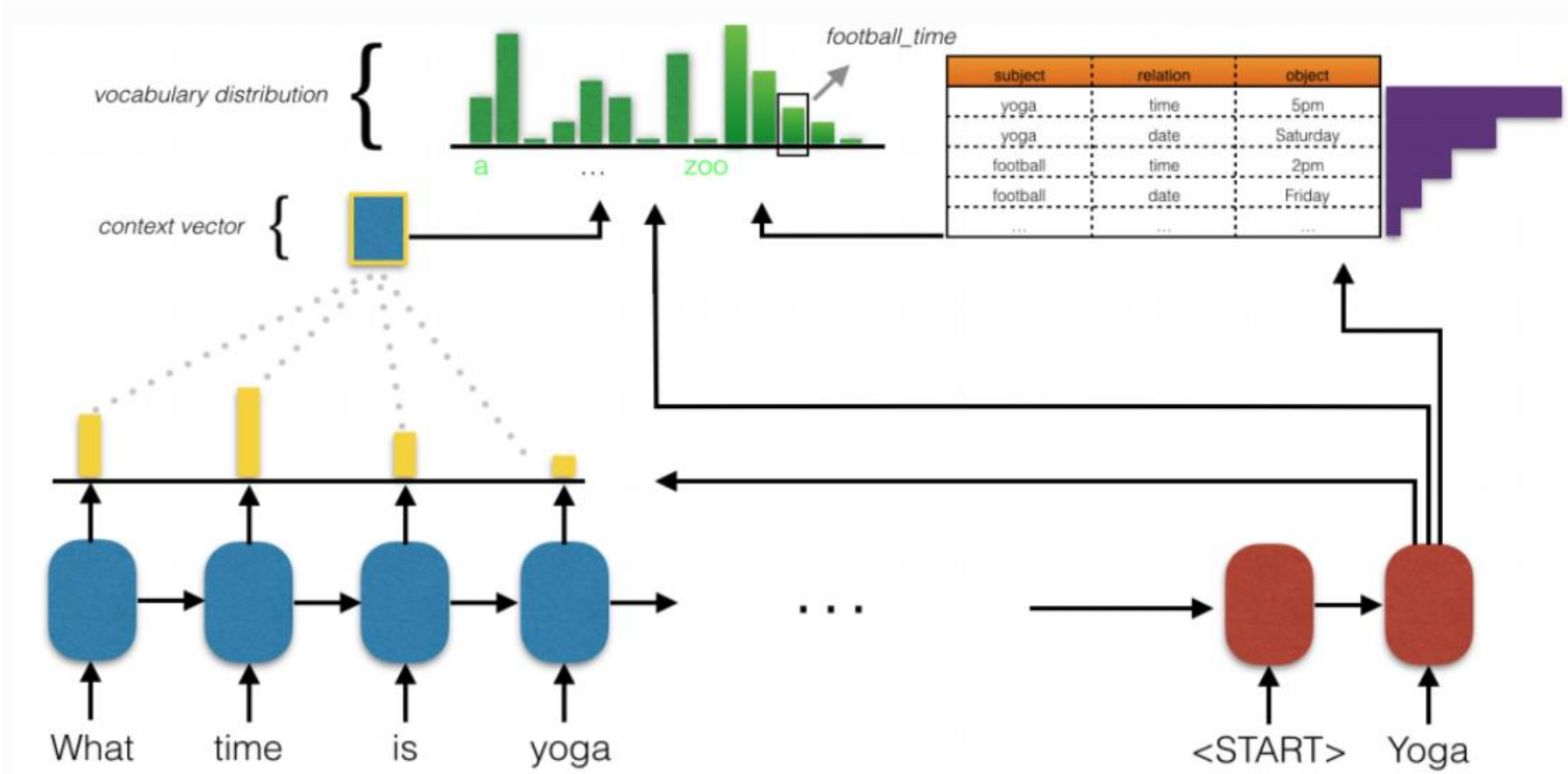
Deeppavlov.ai - возможности

- Определение намерения, тональности
- Морфологическая разметка (11 языков, в т.ч. русский)!
- Определение сущностей (имена, места, организации, даты, количества) с быстрым стартом
- Семантическое ранжирование!
- Коррекция правописания
- Ответ на вопрос цитатой из контекста

Deeppavlov.ai – выполнение текстовых команд



Deeppavlov.ai – ответ фактами из базы знаний



Чатботы – элементы

Морфологический словарь?

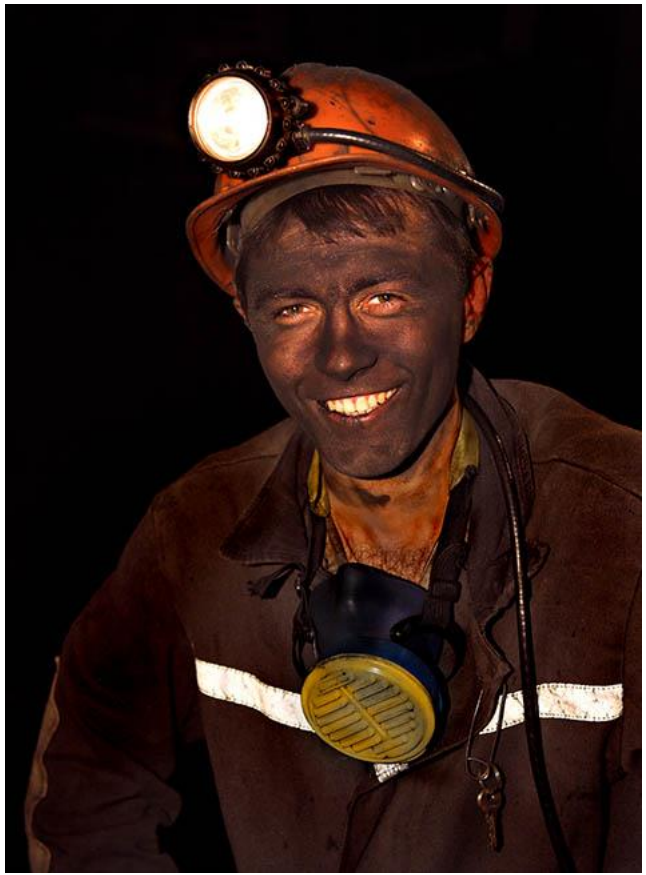
Словарь синонимов?

Семантический граф?

Гипонимы/гиперонимы

Устранение неоднозначностей…

Национальный корпус русского языка…

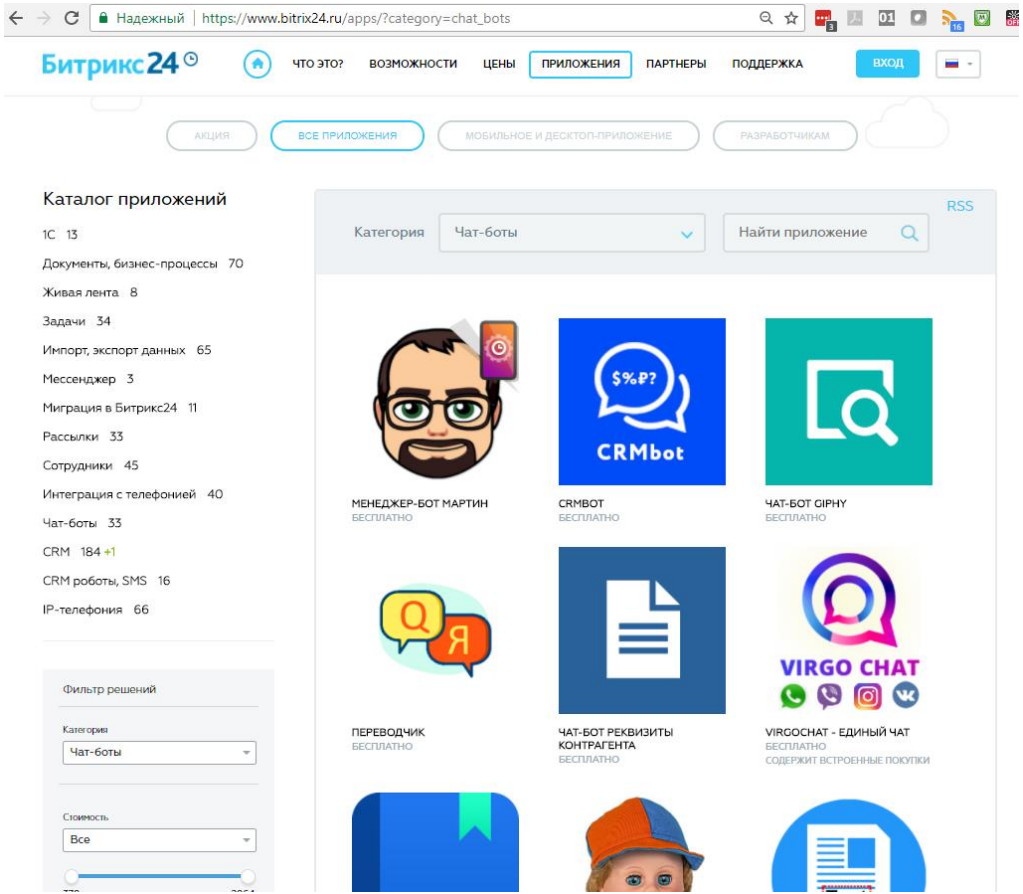


Чатботы – кейсы применения

- Узкая специализация: пиццерия, магазин, справка (ИНН), база знаний, распознавание изображений
- Сбор фактов и выполнение действия (заказ пиццы, билета)
- «Точечное» машинное обучение: анализ тональности, классификация, вычленение фактов

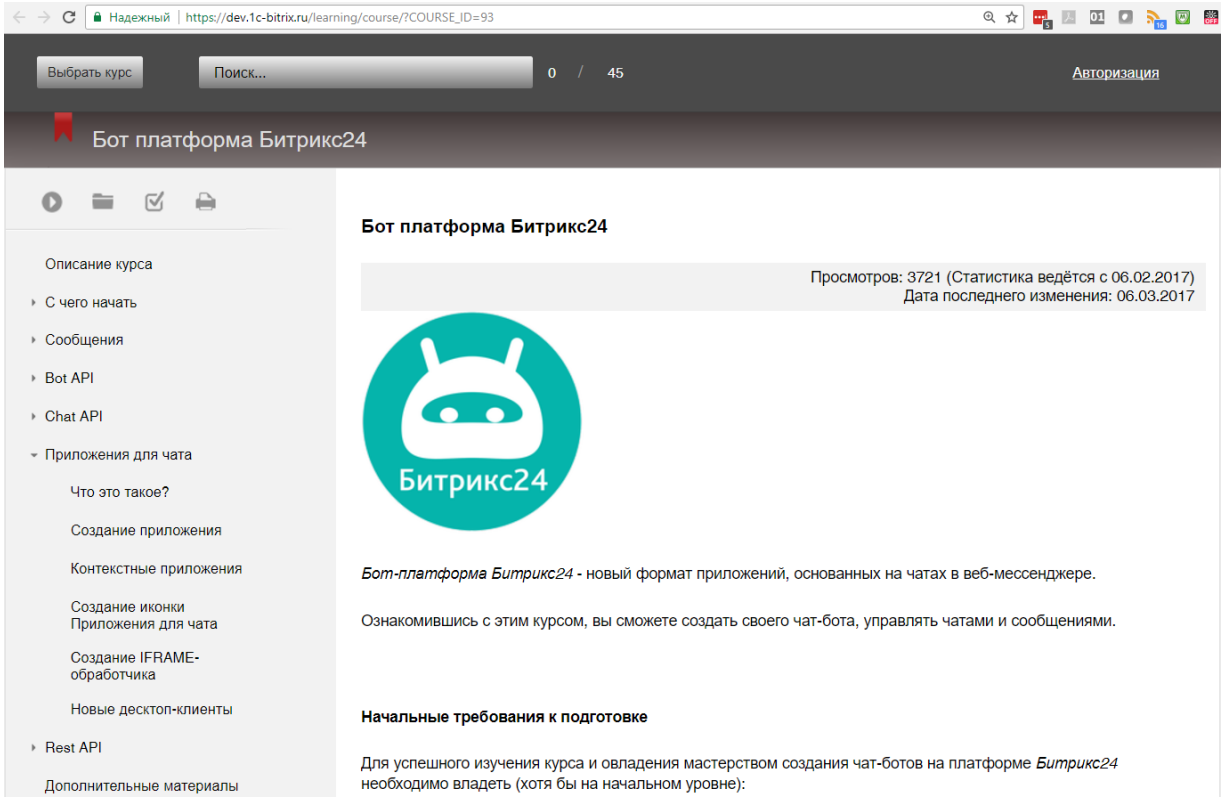
Чатботы – кейсы на Битрикс24

- Можно использовать в открытых линиях в >100к компаний Битрикс24
- Можно использовать в CRM
- Можно использовать в мессенджере
- Современная, удобная платформа для интеграции



Бот плафторма Битрикс24

- Можно писать на любом языке
- Неплохая документация с примерами
- Есть заготовка на PHP
- Веб-хуки



Наши эксперименты

Наши эксперименты:

Ф.М. Достоевский, "Преступление и наказание“

Число слоев сети: 2
Число нейронов в каждом слое: 400
Коэффициент встряхивания "мозгов" (dropout): чуть больше единицы
Память сети: 50 символов назад
Число параметров, которые мы учим - меньше миллиона.

> Старуха, – потому что вы положим, и подумали, как только разве не знаю, как удержать ее. А ведь это не так! Он был в подворотню во все время после него вот и все время совершенно верою и задумчиво. Подошел он к Свидригайлову и почти выскочил вдруг почувствовал, что весь проснулся на него почти в последних порядочную свою подробностей и поскорей почувствовал вопросительно.

> – Не знаю, боже мой! Так вы завтра же объяснить?

> – Подлец! – прошептала Дуня, обернулась к нему.

> – Не знаю, Соня, пожалуйста, – прошептала Дуня, оборачиваясь к нему.

> – А вот ведь только теперь не помню?

> – Покойники подлец! – прошептала Дуня, обернулась к нему.

> – Не возьму нельзя ли к Соне.

> – А ведь я пришел тебе верю. Вы изволили вы присуждают?.. Так я и не знаю чем я показалось, что вы меня от него не подумаете?

Наши эксперименты

Наши эксперименты:

Л.Н. Толстой, "Война и мир"

Число слоев сети: 3
Число нейронов в каждом слое: 400
Коэффициент встряхивания "мозгов" (dropout): чуть больше единицы
Память сети: 150 символов назад
Число параметров, которые мы учим - несколько миллионов

> - Ежели бы мы мог видеть меня в государство, а ему нужно забыть.

> - Andre, je ne partont, j'ai la vaintu qui qu'on russe profende, [Пускай граф и капитан Полковник сказал, что надо делать Кутузовского французского ученица, - говорил он и всегда говорил:

> - Вы не шуби, про положение, - сказал Тихон. - Маменька, господа, - повторил бы насильно вошедший за сеном, - я тебе говог'ю, - сказал Наполеон, так же как и он, взяв свою голову через город, чтобы одно неловко. Старик вздрогнул рукой с штабными голосами, которые стала после полного генерала, указывал он белый длинной улыбкой, стоя в гостиной от одного зуба. Пьер был не при для него теплому. Он был между разговором и невольно присел сепеть; но сосредоточил войско Пьера. Прежде стоит показать что-то. Поворотливый мальчик черного и княжен и говорил об этом, неясное, уверенное и старались изволить все, не чувствуя скорее свое места. Жюли проводила от гибни и довольный, не пытьма, развшед, смеялся; но их было силы и разноречивые своими или мыслью, которые были сделаны из победа, а в доме стояло бравный последние пленных произведенных крестов и деревни. Разговор и старательно сосредоточил то, что он скажет, он не отвечал ей больное с силами и упреков, не переставая помешался в одной минут Петя, ожидая всего того, что сказал при ней и сказал это, прежде их нет; но на этом горы проявилось вперед, в то время как Даву отдавал те слова, сказав очень хорошо, что почти входила в Брюнн и быстро перебирались от усталого, как кто движением голос Земя, которым любил ограничен в эту минуту. В первое время оделялся с ним и велел княжна Марья решительная глаза с большой и гордыми правительством человека и видела повинойствия.

Наши эксперименты

Наши эксперименты:

Код ядра Битрикс

3-х слойная сеть, размер слоя: 400 нейронов, несколько миллионов параметров, память: 150 символов назад, обучение - ночь

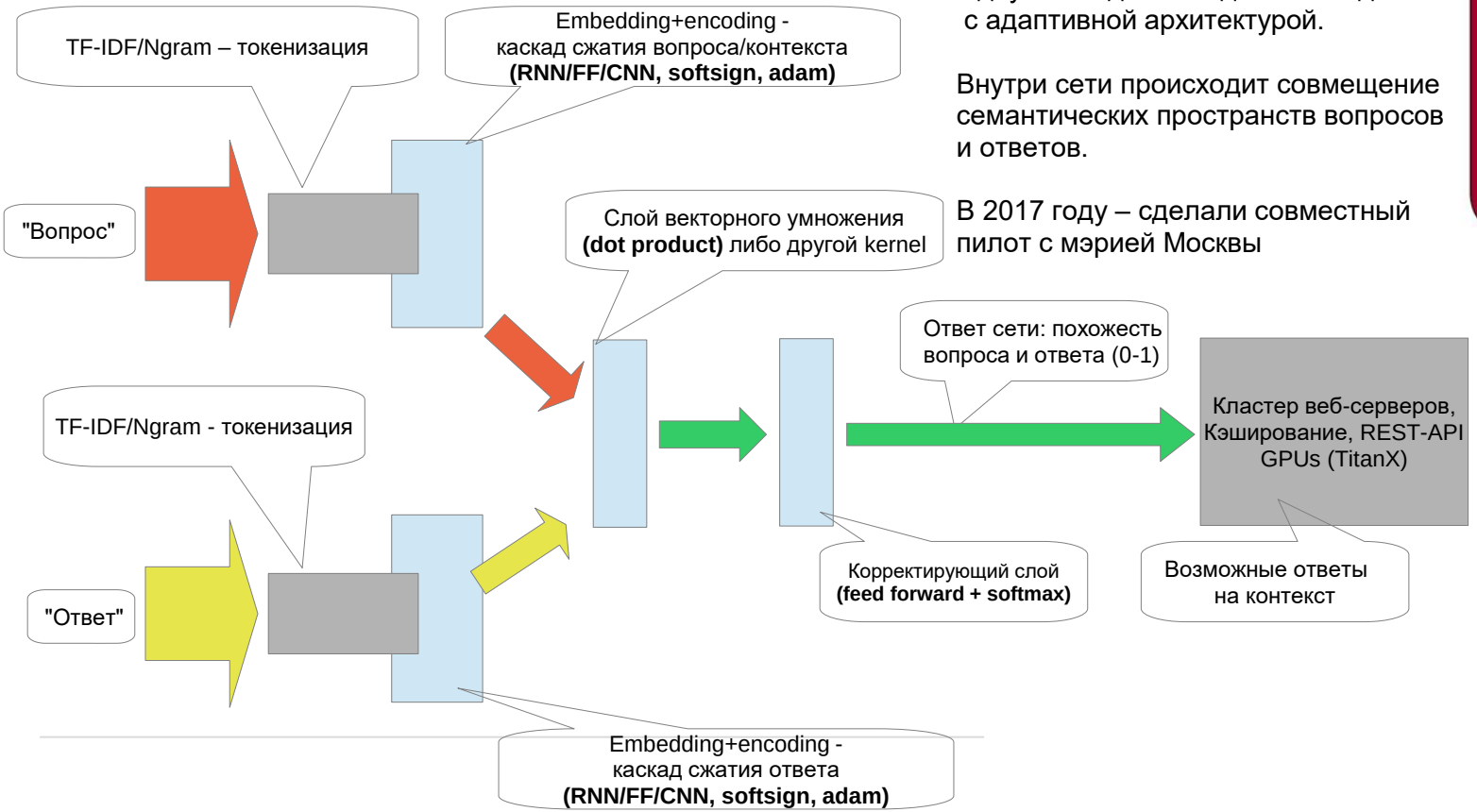
```
function GetGroupsCache($SITE_ID, $site_groups=false)
{
        $_SESSION["SEARCH"] = $arGroups;
        global $USER_FIELD_MANAGER, $USER_AGENT;
        $groups = Array();
        foreach($GLOBALS["SORT"] as $group_id)
        {
                $group_by_id = htmlspecialcharsbx($group_id);
                if ($group_by==".group")
                        $group_id = $group_id;
        }
}

//returns we log old group leave
if(is_array($GROUP_CODE_COUNT))
        $CACHE_MANAGER->Set("USER_CAL_EDIT", $zr);

//Check ID in Rating
if($CURRENT
|| $group_by==0 && check_bitrix_sessid())
{
        COption::SetOptionString('socialnetwork', 'email_provider', 'Y');
        InitBVar($_REQUEST['admin_save_relation']);

        if($group['START_COMMAND_TEXT']=='N')
        {
                $CURRENT_USER_RIGHT = $APPLICATION->GetGroupRight("statistic");
                $discount_value = intval($GLOBALS["USER"]->GetID());
                $runtime->Update($groupId, array("delete filter" => true));
```

«Нейробот»

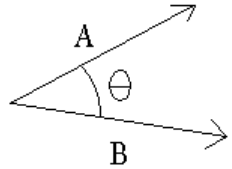
Глубокая нейронная сеть
с двумя входами и одним выходом
с адаптивной архитектурой.

Внутри сети происходит совмещение семантических пространств вопросов и ответов.

В 2017 году – сделали совместный пилот с мэрией Москвы



$$A \cdot B = |A|\,|B| \cos\theta$$

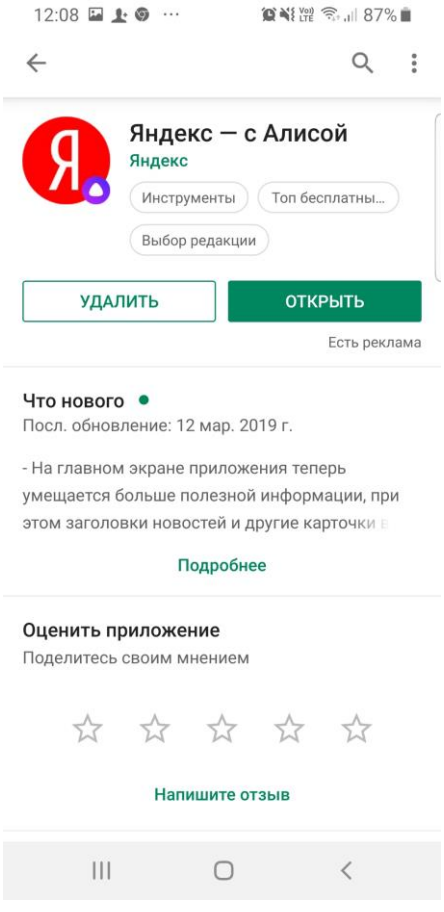


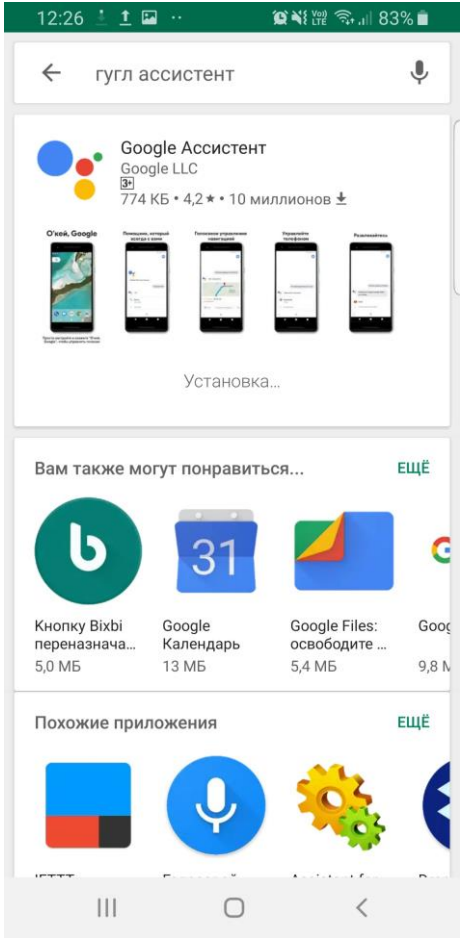
Голосовое управление

Голос вместо текста

- Распознавание речи через нейросеть
- Понимание неоднозначностей
- Яндекс, Google, Amazon

Пример софта для управления





Возможности - Сообщения

Для отправки сообщений

Отправить сообщение.
Отправь сообщение Сергею.
Отправь Сергею сообщение.
Отправь сообщение Сергею Иванову.
Отправить сообщение привет.
Отправить сообщение Сергею привет.
Отправить сообщение Сергею Иванову привет.
Отправить Сергею Иванову сообщение привет.
Напиши сообщение Насте.
Напиши сообщение Насте привет, как дела?
Напиши сообщение Насте с текстом привет, как дела?
Новое сообщение Антону.

Возможности - События

Для создания событий календаря

Добавь событие в календарь.
Добавь событие в календарь на 12 декабря.
Добавь событие в календарь на завтра.
Добавь событие в календарь на завтра с заголовком Маркетинг планерка.
Запланируй встречу на завтра с 14 до 15.
Новая встреча в пятницу после обеда.
Добавь в календарь совещание на завтра на весь день.
Запланируй встречу на 12 декабря на весь день и назови ее стратегическая сессия.
Поставь встречу на следующий четверг 19:00.
Запланируй событие с 12 декабря 10:00 по 15 декабря 15:00.
Запланируй встречу завтра с 11 часов до 13 часов.
Создай собрание с текстом Обсуждаем планы в следующий вторник с 11 часов до 15 часов.
Запланируй встречу сегодня с 11:00 до 13:00.
Запланируй встречу с 12 декабря в 9 часов 31 минуту по 14 декабря 10 часов 19 минут вечера.

Возможности - Задачи

Для создания задач

Создай задачу.
Создать задачу.
Поставь задачу.
Поставить задачу.
Новая задача на завтра.
Поставь задачу на Сергея.
Поставь задачу на Сергея Иванова.
Поставь Сергею задача на завтра после обеда.
Поставь задачу с текстом отчет маркетинга.
Поставь задачу с текстом отчет маркетинга на 12 апреля 2012 года.
Поставь задачу на следующий вторник в 13:15.
Поставь задачу на следующий вторник в 13:15 на Сергея Иванова.
Поставь задачу на Сергея Иванова на следующий вторник в 13:15.
Поставь на Сергея Иванова задачу на следующий вторник в 13:15.
Поставь задачу на Иванова с заголовком привет на 12 декабря в 9 вечера.
Добавь Сергею Иванову задачу на послезавтра и назови ее привет.
Добавь задачу с ответственным Сергеем Ивановым
сроком следующий вторник и текстом привет.
Добавь задачу где ответственный Сергей Иванов со сроком
на следующий вторник и текстом привет.

Подключение – «Алиса» на портале Битрикс24

Подключить голосовой помощник Алиса

1 Активируйте свою колонку или запустите приложение

Навык Битрикс24 Ассистент можно активировать в любом продукте Яндекса, в который встроена Алиса — в Яндекс.Станции, Браузере, Алисе для Windows и т.д.

2 Произнесите: Алиса, запусти навык Битрикс24

3 Дождитесь, когда Алиса попросит у вас код

После запуска навыка Алиса попросит вас ввести код авторизации.

4 Произнесите код, который был сгенерирован Битрикс24

[redacted] [redacted] 5 5 2

ВСЁ ПОЛУЧИЛОСЬ

Подключение – «Google Ассистент» на портале Битрикс24

Подключить голосовой помощник Google Ассистент

1 Активируйте свою колонку или запустите приложение

Приложение Битрикс24 Ассистент можно запустить в любом продукте, где встроен Google Ассистент — в Google Home, в Android и т.д.

2 Произнесите: OK Google, поговорить с Битрикс24

3 Дождитесь, когда Google Ассистент попросит у вас код

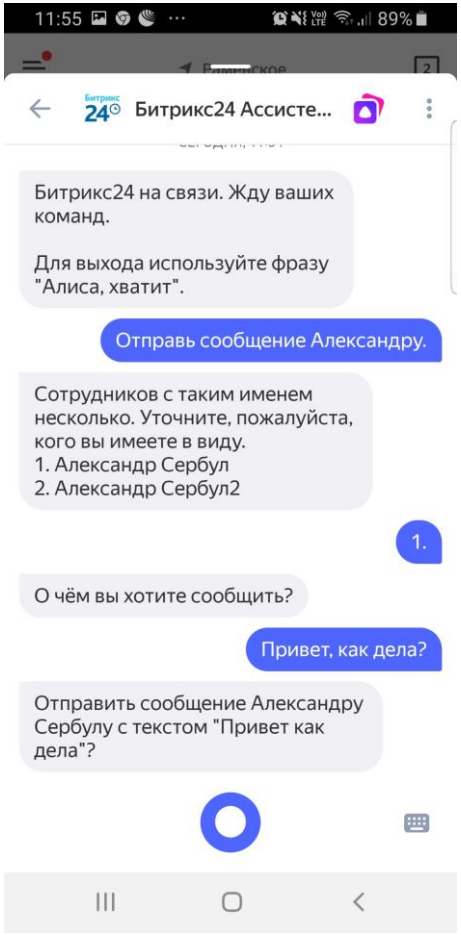
После запуска приложения Google Ассистент попросит вас ввести код авторизации.

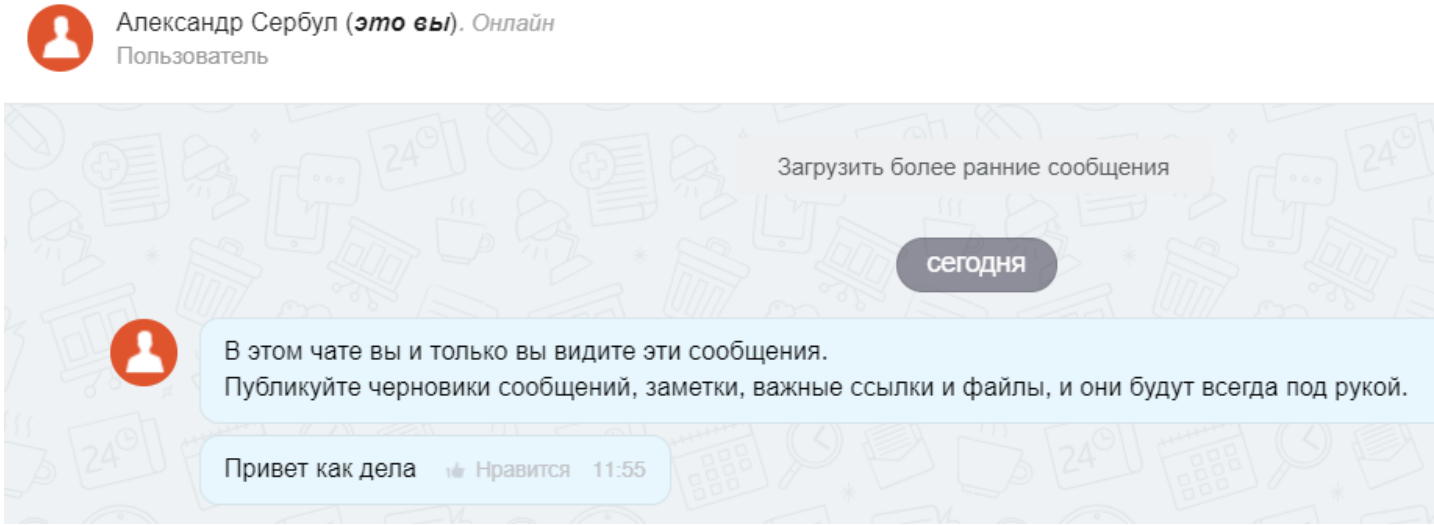
4 Произнесите код, который был сгенерирован Битрикс24

[redacted] [redacted] 7 4 9

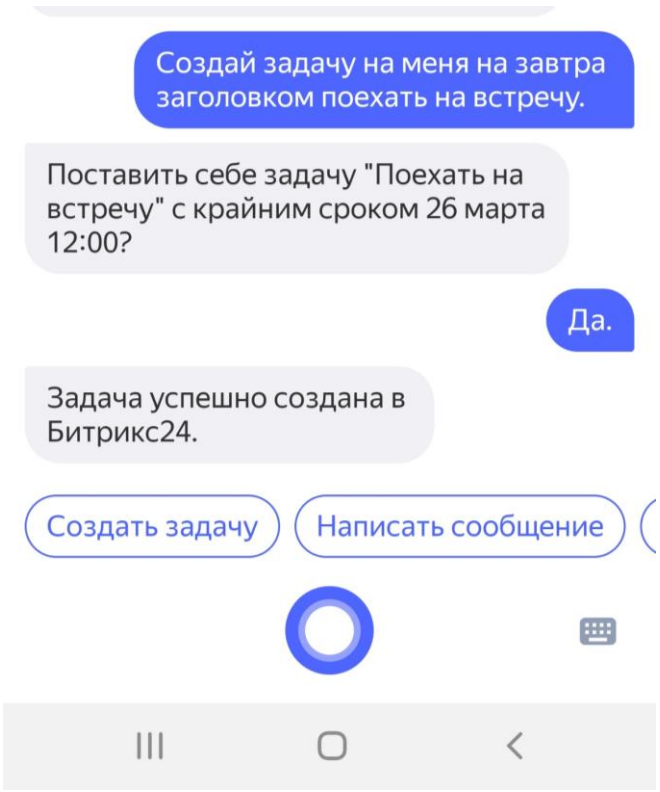
ВСЁ ПОЛУЧИЛОСЬ

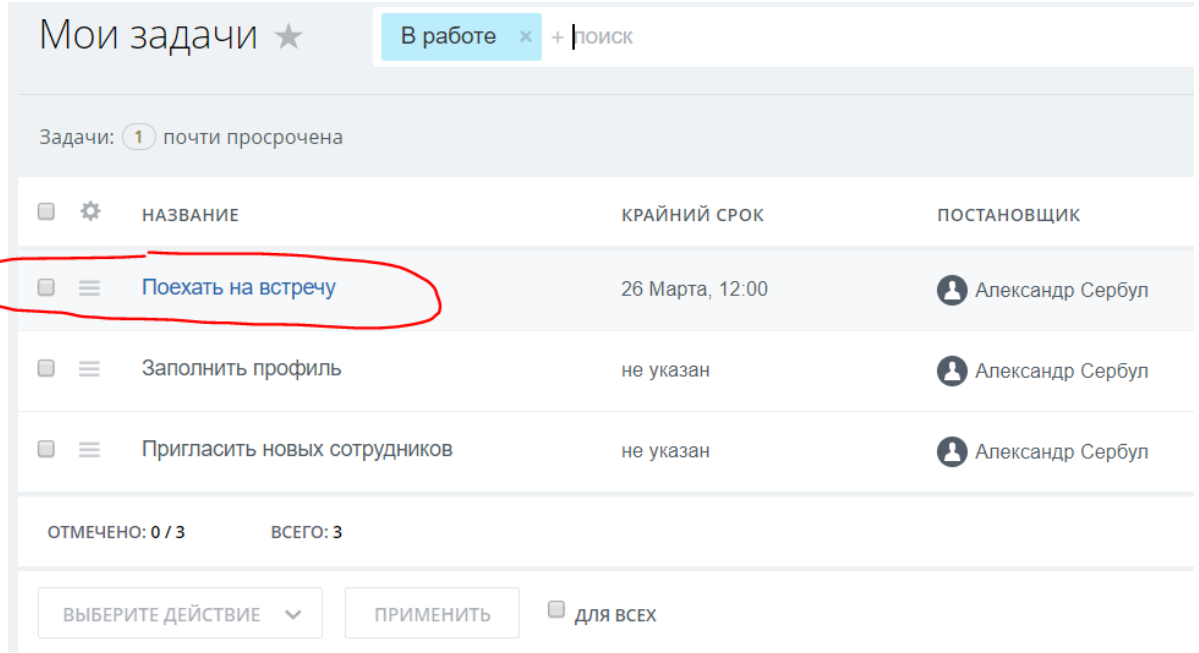
«Алиса», отправка сообщения на портале Битрикс24



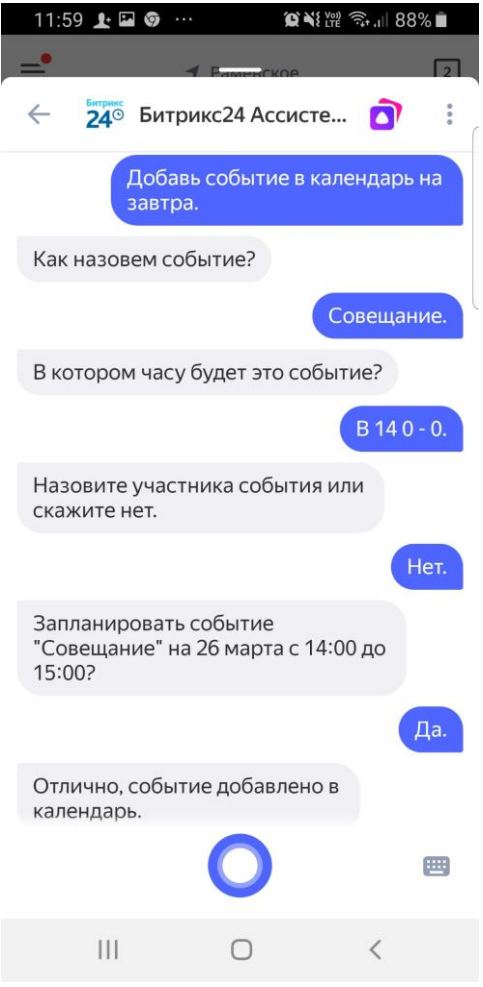


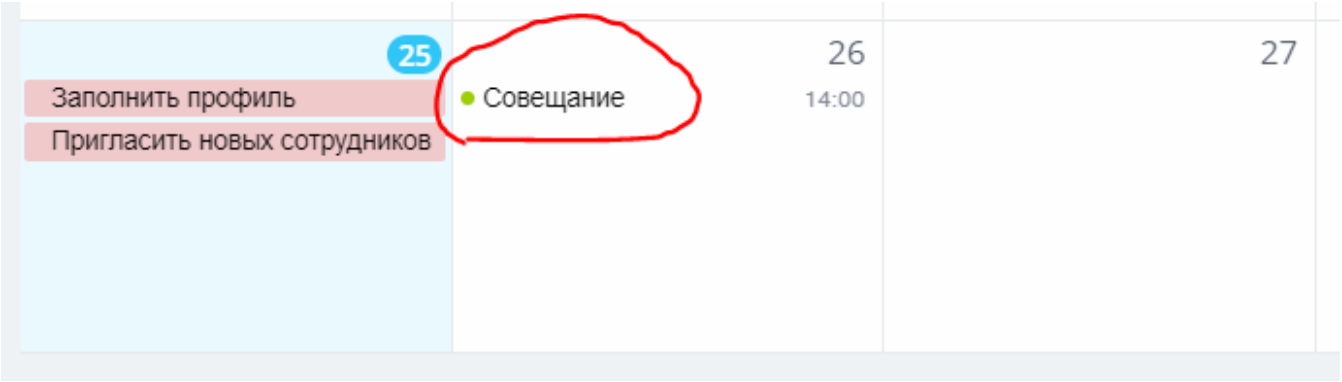
«Алиса», создание Задачи на портале Битрикс24



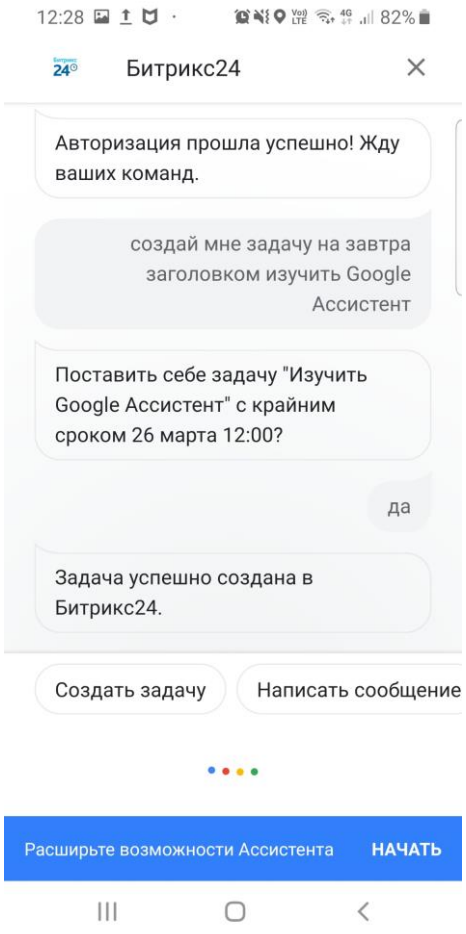


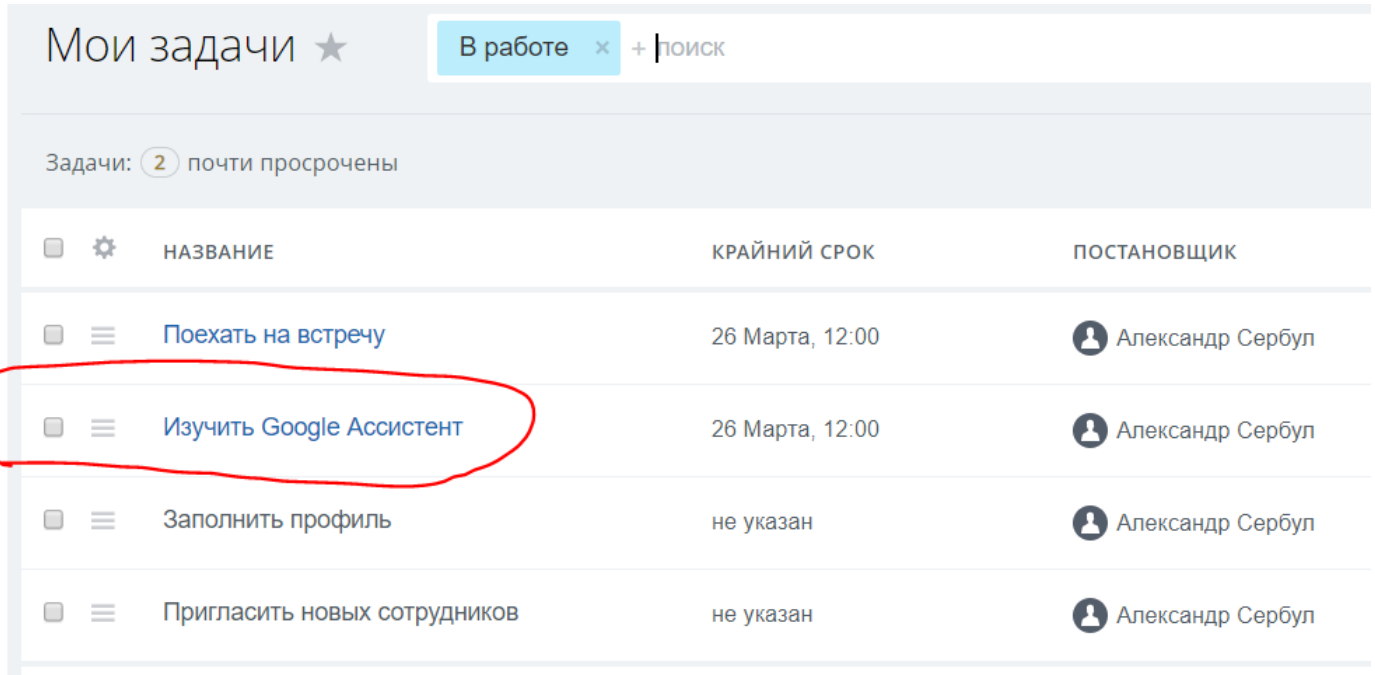
«Алиса», создание События на портале Битрикс24





«Google Ассистент», создание События на портале Битрикс24





Где брать людей в команду?

- Акселераторы, хакатоны
- Физтех-акселератор (pha.vc)
- Сообщество «OpenDataScience» (ods.ai)
- aione.world, «ScienceGuide»

Синергия, синергия, синергия…

Где брать людей в команду?

- Бигдата: хорошие программисты и опытные сисадмины – 1 штука на проект
- Создание/тюнинг моделей: физматы – 1 штука на отдел
- Product owner с обновленным мозгом – 1 штука на проект(ы)
- Менеджеры – 1024 килограмм ☺
- python, java, unix, spark, scala, julia

Спасибо за внимание!
Вопросы?

Александр Сербул
@AlexSerbul
Alexandr Serbul
serbul@1c-bitrix.ru