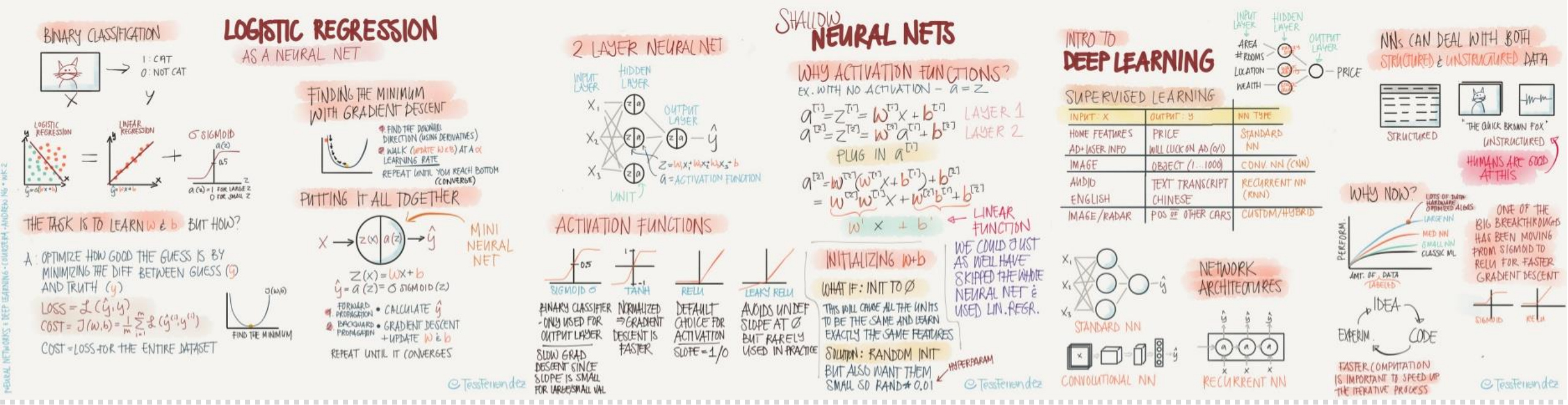


Natural Language Processing

в банковской сфере

Семинар МШЭ МГУ, 11.10.2021

Диана Котерева, Сергей Афанасьев

Ренессанс Кредит

МГУ Московская школа экономики МШЭ

Коротко о нас

Диана Котерева

Head of R&D

КБ «Ренессанс Кредит»

Сергей Афанасьев

Vice-President
Chief Data Scientist

КБ «Ренессанс Кредит»

Банк «Ренессанс Кредит»

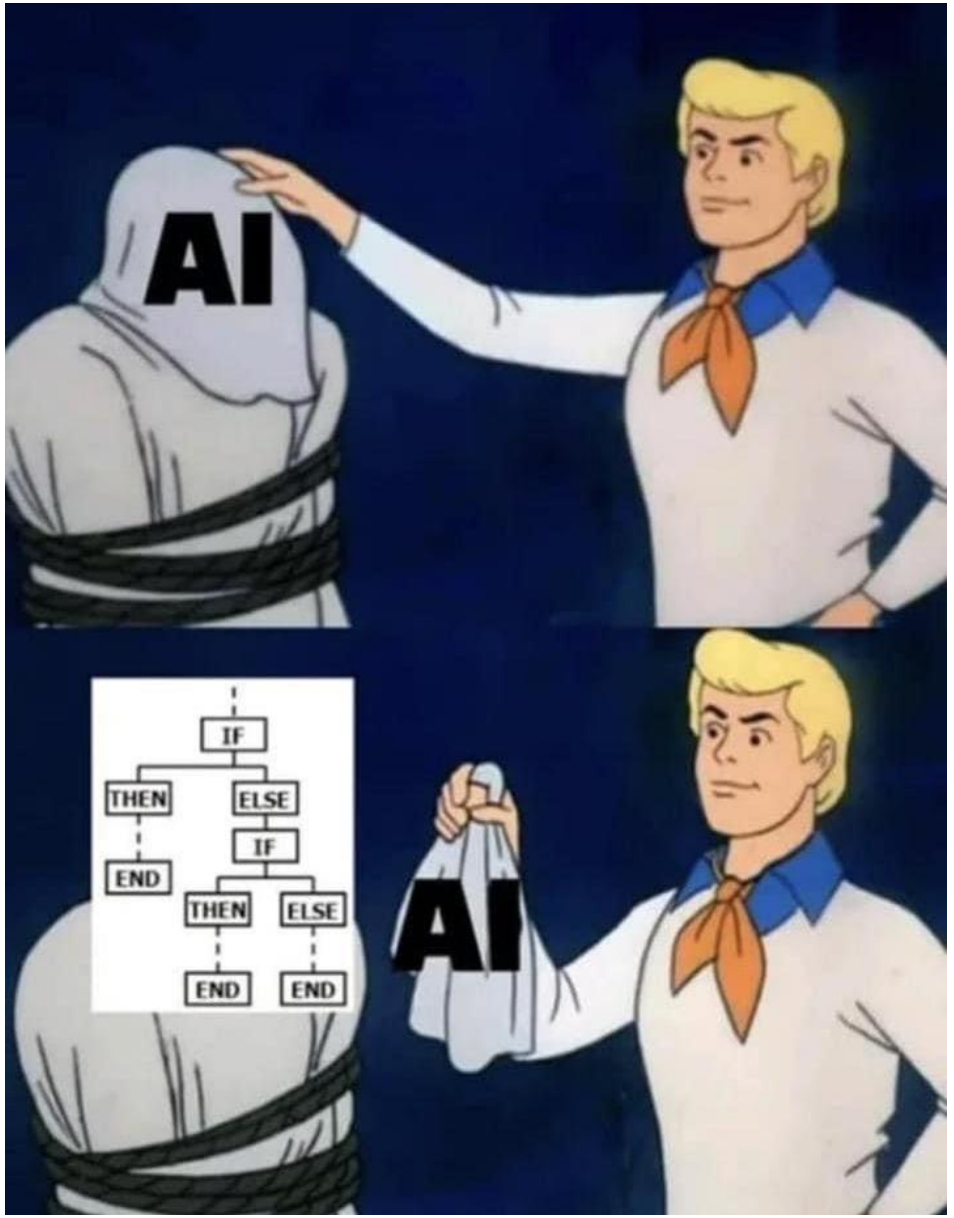
75 млн
заявок

13 млн
клиентов

9 млн
кредитов

– Как отличить **Machine Learning** от AI?

– Если написано на Python – это **Machine Learning**, а если на PowerPoint, то AI.



Data Science в банковской сфере

	Risk, Anti-fraud	Collection	CRM	CS, CC, TM	IT, IT Security	Other
Classical Machine Learning	**Credit Scoring** - Application scoring - Behavioral scoring - Combine scoring **Anti-fraud Models** - Внутр/внеш fraud - Транзакционный fraud **IFRS-9/Basel III**	**Collection Scoring** - Entry/Self_cure модели - DPD модели - Модели Recovery Rate	**Recommender System** - PTB модели **Attrition model** - Модели оттока клиентов (Churn prediction) **Targeted Advertising** - Таргетирование рекламы		**Biometrics** - Keystroke Dynamics & Mouse/Touch Movements - Device Print для ИБ/МБ **Error Detection** - Выявление сбоев в системах Банка	**AML** - AML правила и модели **Time Normalization** - Оптимальная загрузка массовых подразделений **Staff Recruitment** - Отбор резюме по ключевым словам
Natural Language Processing	**Feature Engineering** - Предикторы на текстовых данных для Скоринга **Voice Recognition** - Аутентификация клиентов по голосу **Fraud Text Mining** - Анализ корп. переписки	**Feature Engineering** - Предикторы на текстовых данных для Collection **Text-To-Speech** - Голосовой коллектор	**Feature Engineering** - Предикторы на текстовых данных для CRM-моделей **Recommender System** - Анализ транзакции с помощью RNN для моделей отклика	**Speech recognition** **CS Text Mining** - Извлечение информации - Анализ тональности - Чат-боты (сайт, ИБ/МБ) **Text-To-Speech** - Синтез речи (Роман-2.0)	**Security Text Mining** - Проверка email-писем на утечку информации и т.п. **Topic Model** - Распределение заявок в Help-Desk по тематикам	**Post Mail Classifier** - Распределение корреспонденции по тематикам **HR Text Mining** - Анализ переписки и выявление проблем в работе сотрудников
Computer Vision	**Photo Biometrics** - Вход в ИБ/МБ - Вход в системы Банка **Object Detection** - Скоринг по фото клиента - Проверка фото партнеров **Spoofing** - Выявление подделок			**OCR (Text recognition)** - Перевод скан.копий документов в текстовый формат.		**Photo Biometrics** - Аутентификация клиентов в ДО/ККО - Пропускные системы для сотрудников

Natural Language Processing (типы задач)

Текстовая аналитика

- Анализ тональности
- Извлечение структурированной информации из текста
- Классификация текстов
- Кластеризация текстов
- Информационный поиск

STT and TTS

- Распознавание речи (Speech-To-Text)
- Синтез речи (Text-To-Speech)

Генерация текста

- Машинный перевод
- Обобщение текста
- Аннотирование текста
- Упрощение текста
- Изменение стилистики текста

Чат-боты

- Разговорные чат-боты
- Вопросно-ответные системы
- Ассистенты (Алиса, Siri)

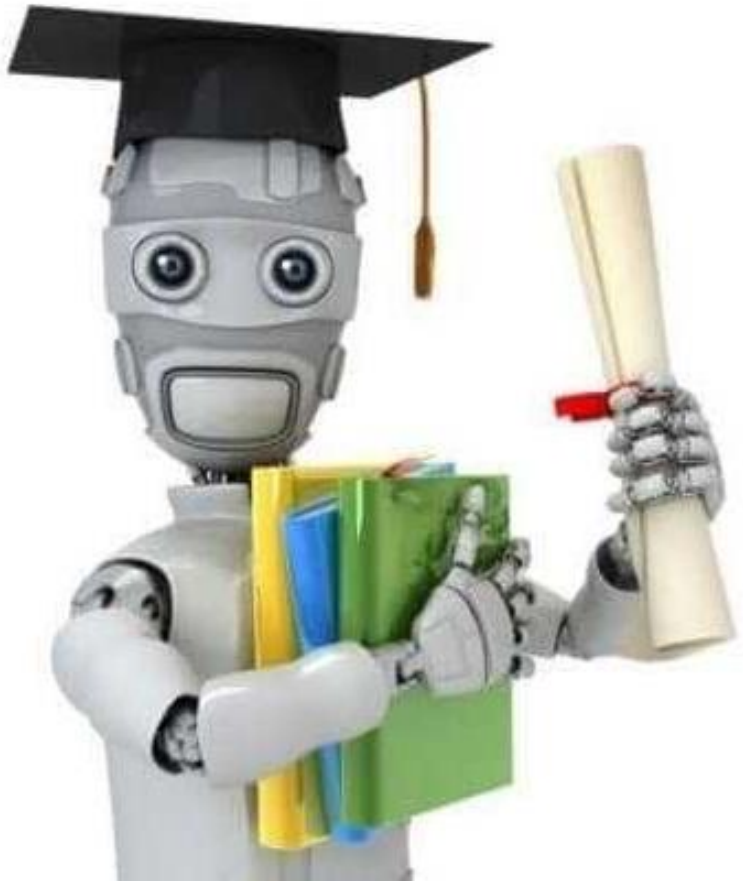
NLP-моделирование в банковской сфере

Котерева Д.М.
Группа моделирования и оперативного анализа
Управление статистического анализа
11 октября 2021 г.

Основные задачи Data Science в банке

МОДЕЛИ

- Risks (scoring)
- CRM
- Collection
- Antifraud
- R&D (NLP, CV, etc)



NLP для подразделений банка

Обогащение банковских моделей

CRM

Обогащение CRM-моделей текстовыми данными

+2,5% GINI

Collection

Обогащение моделей взыскания текстами

Risks

Обогащение скоркарт текстовыми данными

Сервисные NLP-задачи

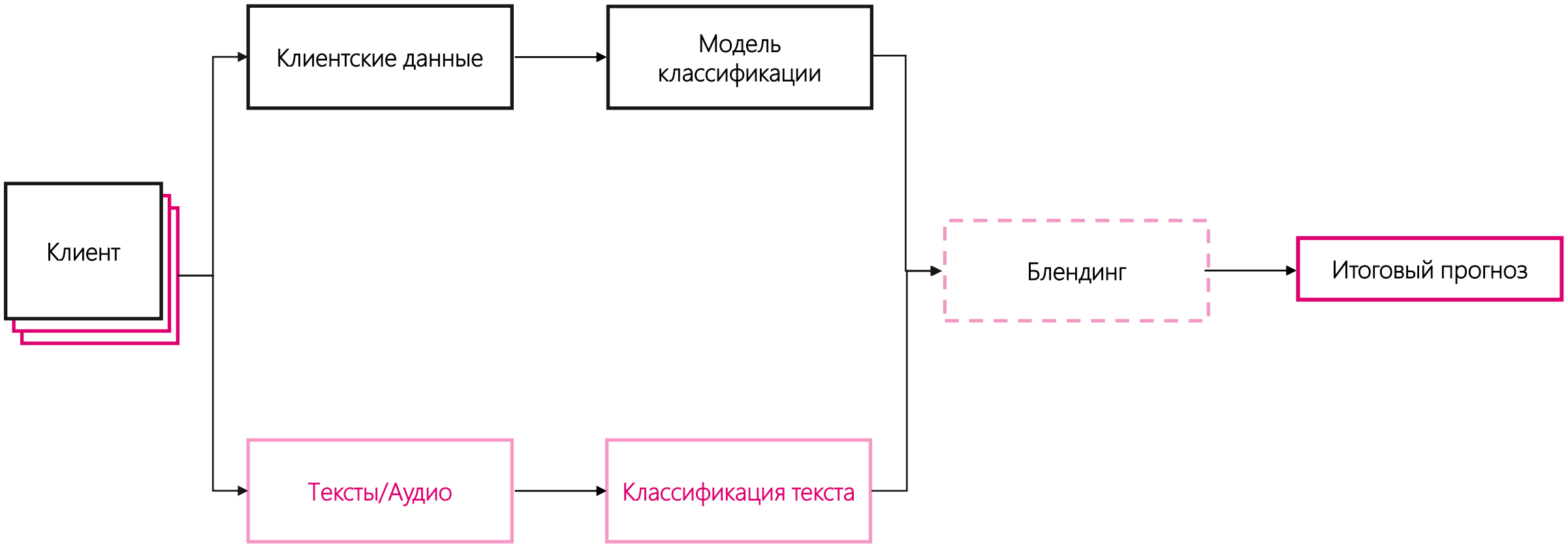
Customer Service

Others

Collection, AML, CRM, Antifraud, IT-Security и др.

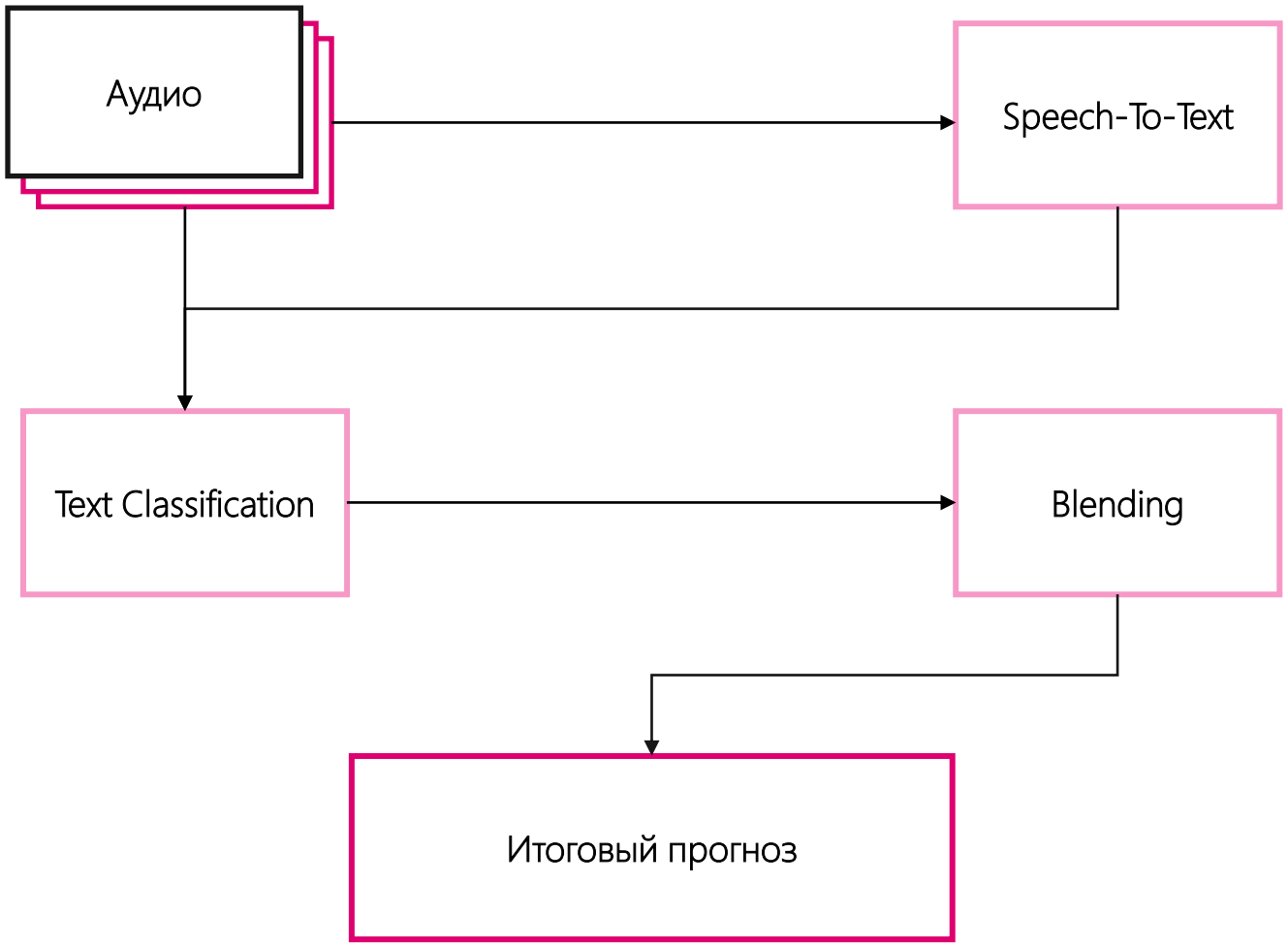
Ренессанс Кредит Банк всевозможностей

Обогащение банковских моделей



Обработка аудио и текстовых данных

Источники:
- Записи звонков
- Тексты (чаты, обращения)



Модели распознавания текста из аудио записи

Обучение классификатора на том же таргете, что и основная модель (Risks, Collection, CRM)

Объединение прогнозов основной и текстовой моделей с помощью блендинга

Ренессанс Кредит Банк всевозможностей

Speech-To-Text

1. Цель модели: Получить оптимальную последовательность слов для заданной последовательности звуков.

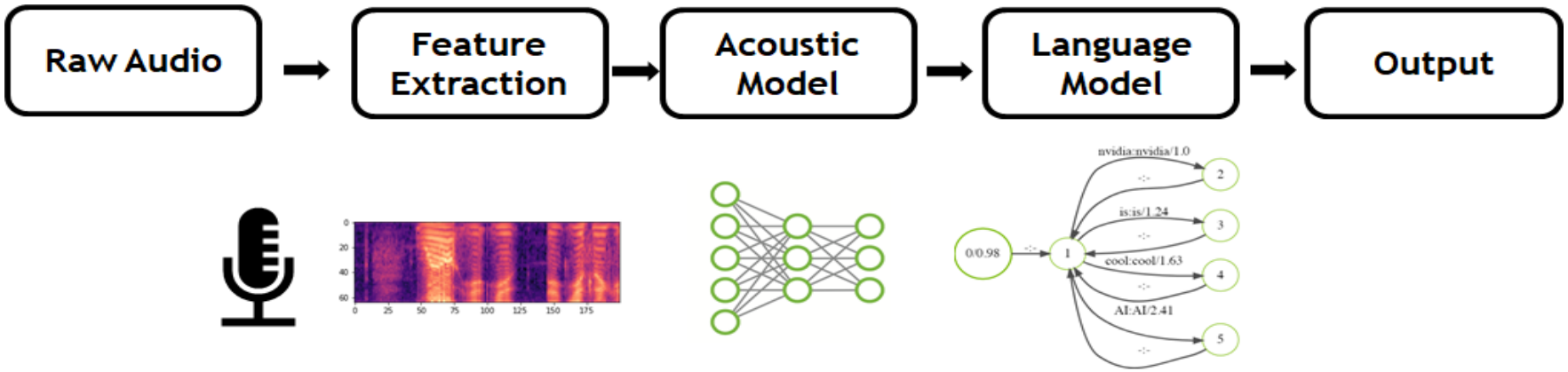
2. Дано: Необработанный аудиосигнал.

3. Процесс обработки:

    3.1. Получение признаков из аудиосигнала (Извлечение MFCC)

    3.2. Получение последовательности текстовых признаков (акустическая модель)

    3.3. Получение последовательности слов (лингвистическая модель)



Извлечение признаков аудиосигнала

Аудио сигнал

↓

Пре-эмфазис (Pre-emphasis)

Подавление шумов

↓

Разбиение сигнала на окна

Увеличение стабильности получаемых признаков

↓

Преобразование Фурье

Получение спектра и периодограммы

↓

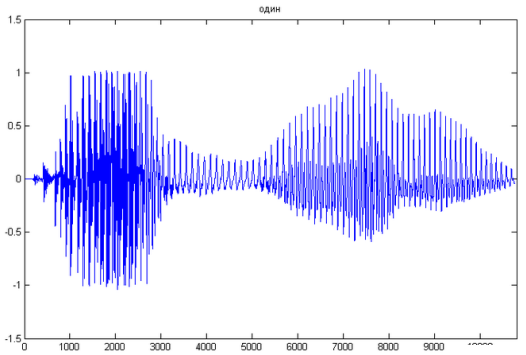
Дефолмация частот по шкале Mel

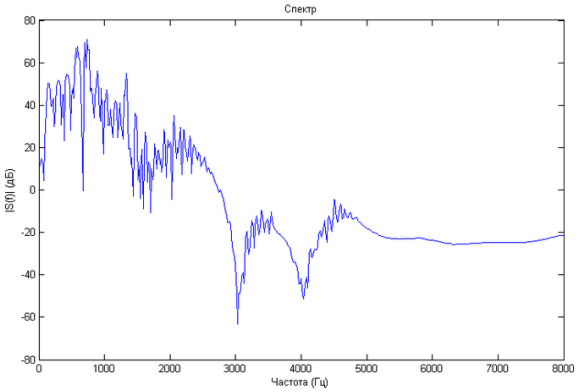
Имитация человеческого слуха

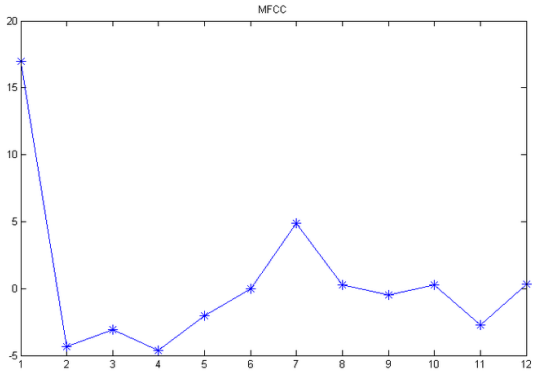
↓

Discrete Cosine Transform (DCT)

Получение энергий

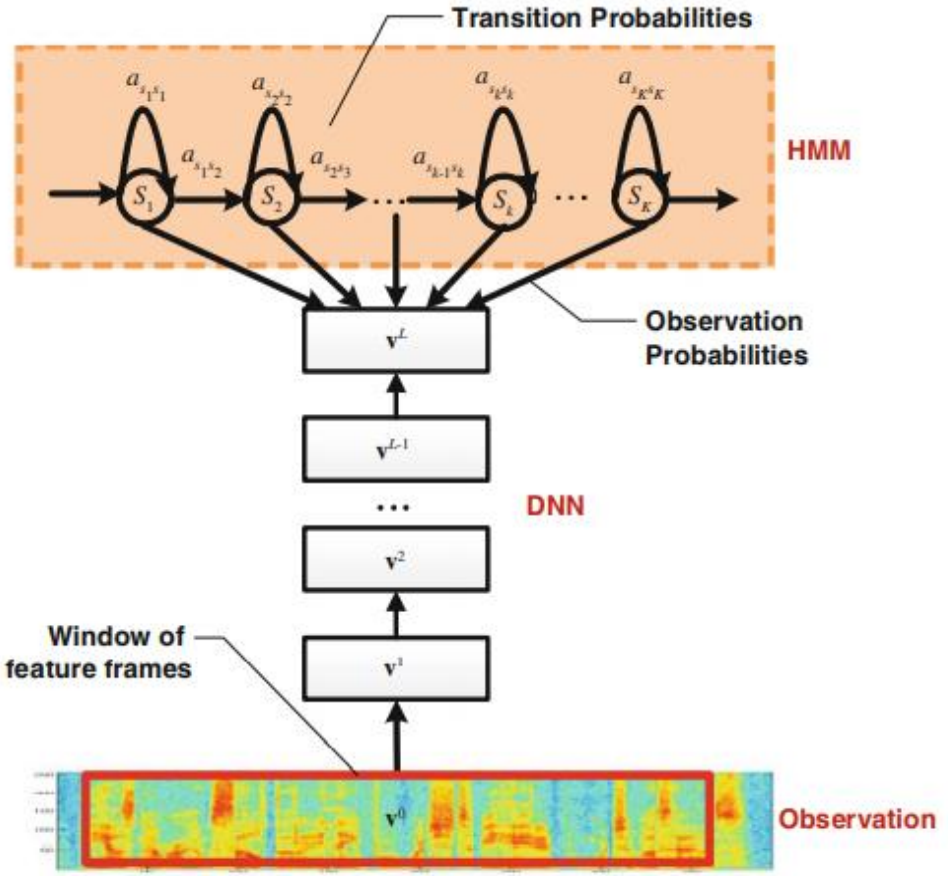






Ренессанс Кредит Банк всевозможностей

Лингвистическая и акустическая модели



$$\hat{w} = armax_w p(w|x) = \frac{argmax_w p(x|w)p(w)}{p(x)} = argmax_w p(x|w)p(w)$$

Акустическая модель

$$p(Acoustic|\mathrm{HMM}State) = \frac{p(\mathrm{HMM}State|Acoustic)p(Acoustic)}{p(\mathrm{HMM}State)}$$

- $p(Acoustic|\mathrm{HMM}State)$ оценивается с помощью HMM
- $p(\mathrm{HMM}State|Acoustic)$ оценивается с помощью глубокой нейронной сети
- $p(\mathrm{HMM}State)$ – априорная вероятность получить каждое состояние, оцененное по тренировочным данным

Лингвистическая модель

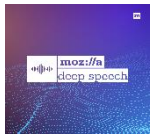
$p(w) = p(w_k|w_{k-1},..,w_1)$- Оценивается только по текстовому корпусу

Ренессанс Кредит Банк всевозможностей

Реализации Speech-To-Text моделей

Kaldi

a. Отдельное использование акустической и лингвистической моделей
b. В качестве лингвистической модели используется CD-HMM-DNN

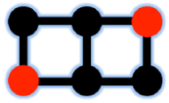


DeepSpeech&DeepSpeech 2

a. Основа модели – RNN слои
b. Единый end-to-end подход

Wav2Letter

a. Основа модели – CNN слои
b. Единый end-to-end подход



EspNet

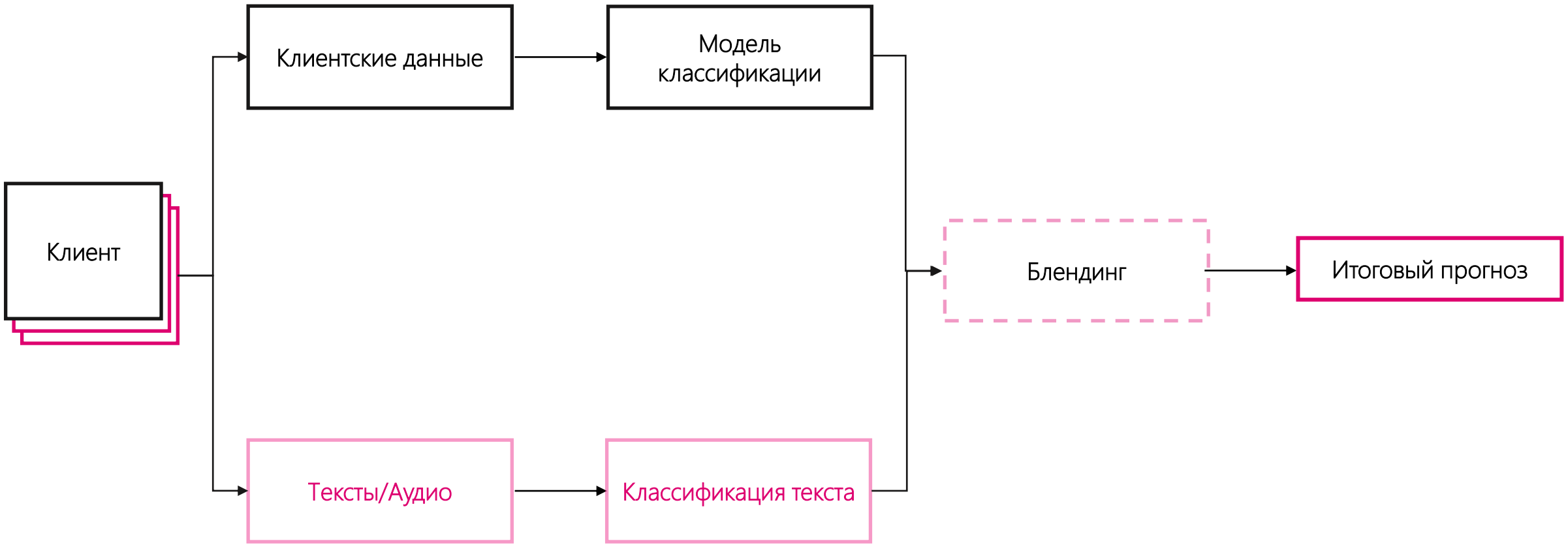
a. Основа модели – Attention
b. Единый end-to-end подход
c. Предобработка сигнала аналогично Kaldi

QuartzNet

a. Основа модели – 1d Channel Separable Convolutions
b. Единый end-to-end подход

Ренессанс Кредит Банк всевозможностей

Обогащение банковских моделей



Обогащение моделей Роботом-Коллектором

Модель: ENTRY

Текстовые данные: пилот с роботом коллектором (аутсорс). Тексты звонков брались до даты скоринга.

	Train	Test	σ
Entry	49,0%	51,2%	
Entry + Robot	53,3%	53,6%	
Δ Gini	**4,4%**	**2,4%**	**0,7%**

Выводы:

По сравнению с Романом диалоги Робота-Коллектора **информативнее** за счет лучшей адаптации робота под речь оператора

Интерпретатор текста (ELI5)

robot здравствовать вы приветствовать автоматизированный помощник «банка ренессанс кредит». Информировать вас, данный разговор записываться фио я мочь услышать?

customer это я

robot мы звонить вы по повод ваш просрочить задолженность перед банк ренессанс кредит по денежный кредиту. На сегодняшний день у вы иметься задолженность в размер сумма_деньги вы готовый оплатить её в течение три дней?

customer три нет на на следующий неделя оплатить

robot фио возврат в график платёж позволять прекратить звонок и другой вид информирование по повод просрочить задолженность с сторона «банка ренессанс кредит». Вы готовый оплатить просрочить задолженность в течение три дней?

customer в три нет на следующий неделя оплатить в понедельник

robot фио обращать ваш внимание, что весь информация о возникновение просрочить задолженность фиксироваться в бюро кредитный история с указание срок и сумма просрочить задолженности. Вы готовый оплатить просрочить задолженность в течение три дней? К сожаление вы не слышный.

customer в понедельник в понедельник.

Примечание: текст диалога лемматизирован для моделирования

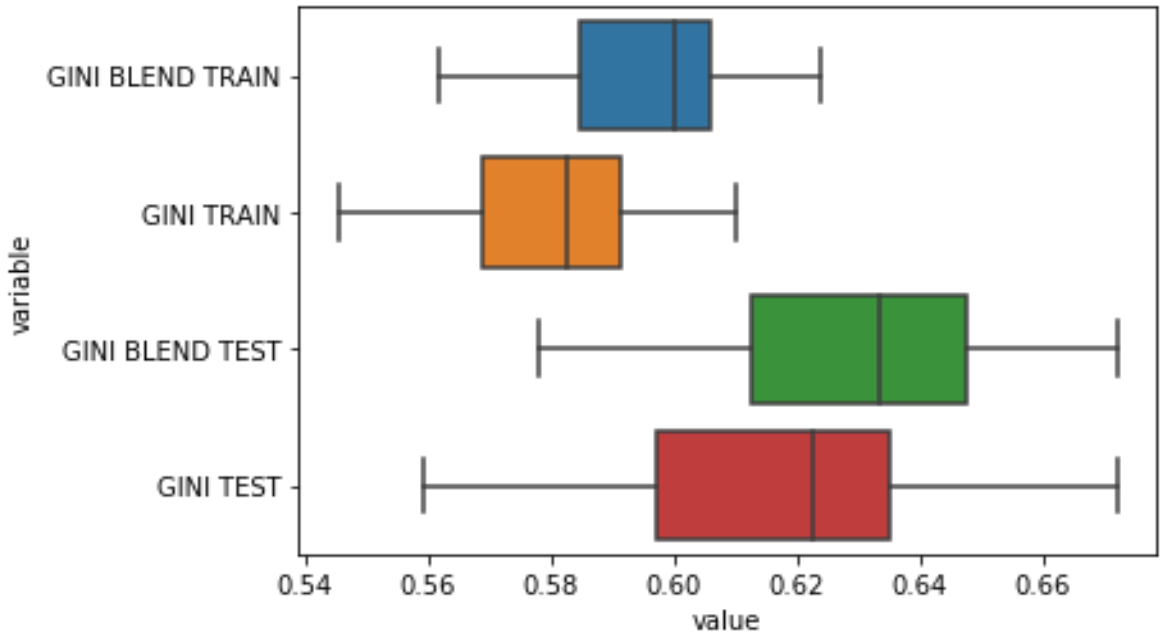
Ренессанс Кредит Банк всевозможностей

Обогащение моделей Outsource Collection звонками

Модель: SW

Текстовые данные: записи outsource звонков Collection в течение года до даты скоринга

Blending Test Sample				
Показатель	Среднее	5%	**Медиана**	95%
GINI_без текстов	61.72%	56.98%	62.25%	66.05%
GINI_с текстами	63.05%	58.64%	63.35%	66.57%
Прирост GINI	**1.33%**	-0.16%	**1.40%**	2.96%



Выводы:

Предыдущие записи контакта клиента с Collection информативны для будущей классификации

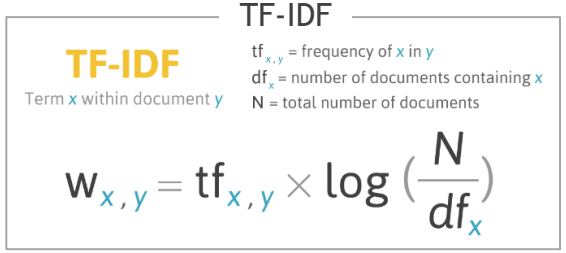
Ренессанс Кредит Банк всевозможностей

Классификация с учетом диалоговой структуры

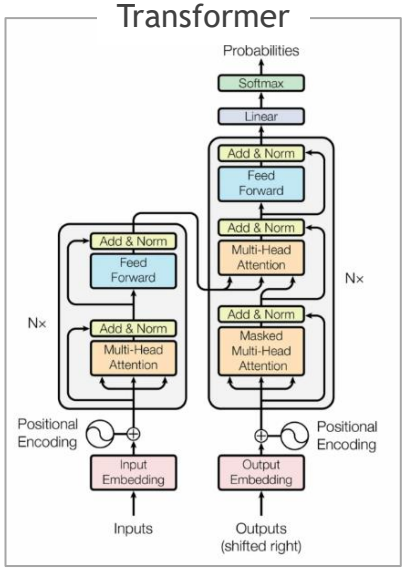
У оператора и клиента разные паттерны поведения => их надо **разделить**!

- Дополнительные обучающие токены, указывающие на актора
  - Transformers
  - RNN ячейки
- Разделение представлений фраз клиента и оператора
  - DialogueRNN
  - Реализации для Transformers могут быть дороги

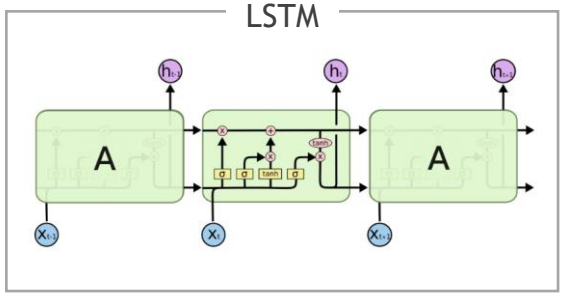
TF-IDF



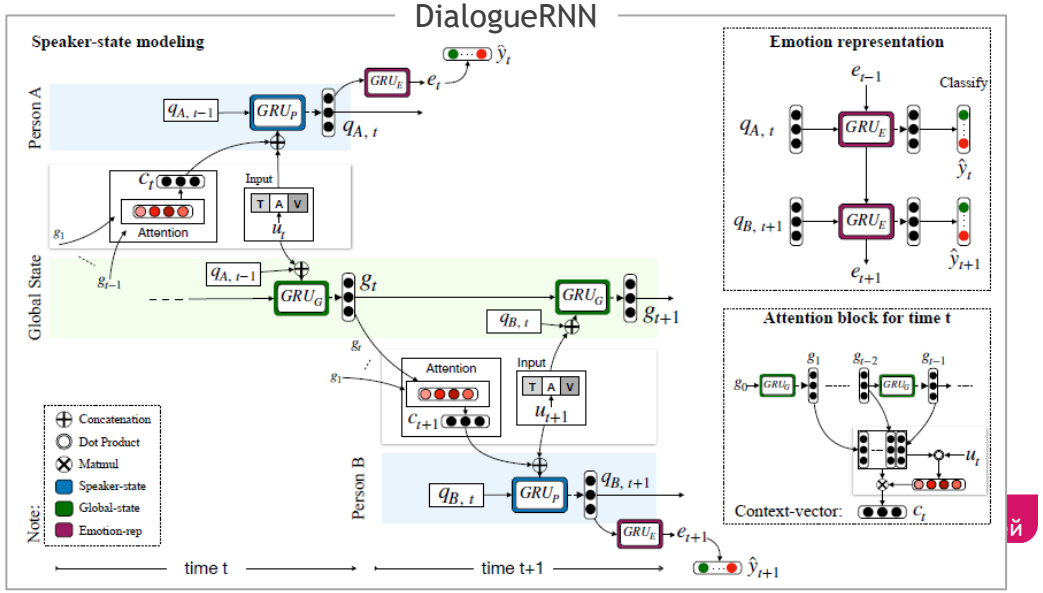
Transformer



LSTM



DialogueRNN



Appendix

Примеры моделей банков из топ-1

Риски

1. Анализ тональности новостных текстов для заданного определения тональности
2. Задача выявления нерациональных сделок
3. WWR при расчете цены производных инструментов
4. Разработка моделей эволюции для макроэкономических параметров, релевантных для заданной области промышленности
5. Задача оптимизации процедуры выставления ограничений на торговый портфель банка
6. Модель для оценки надежности сложной системы
7. Задача поиска разладки в условиях ограниченного набора данных
8. Совместное симулирование рыночных риск-факторов при помощи копул.
9. Моделирование рыночных риск-факторов методом Монте-Карло с использованием стохастических процессов со скачками.

Риски ликвидности

1. Кластеризация банков по поведению клиентов и моделям ведения бизнеса
2. Разработка алгоритма определения банков с повышенным спросом на ликвидность
3. Моделирование рынка вкладов физических лиц в валюте
4. Моделирование рынка депозитов юридических лиц в валюте
5. Разработка рекомендательной системы по оптимальным периодам первичного размещения облигаций
6. Определение справедливой ставки привлечения по вкладам физических лиц для различных банков
7. Определение оптимальной структуры розничного портфеля
8. Разработка неценовых инструментов управления клиентским балансом банка на рынке с жестким государственным регулированием
9. Оценка влияния развития sharing economy на потребительский спрос и вероятность реализации кризиса перепроизводства

Бизнес

1. Извлечение данных из комментариев к объявлениям
2. Построение маршрута с мин. затратами и макс. Эффектом
3. Оценка качества и потенциала бизнеса клиента банка с применением анализа неструктурированных текстов
4. Предсказание совершения покупки физлицом на основе его транзакций
5. Определение поведенческой модели пользователя на основе его трат
6. Определение контекста с предсказанием следующих действий пользователя
7. Оценка качества и потенциала бизнеса клиента банка с применением NLP инструмента
8. Извлечение ценовых параметров из объявлений о продаже/аренде недвижимости
9. Textual Entailment в задаче классификации документов
10. Анализ эффективного использования ценовых участков города

ИТ

1. Оптимизация хранения текстовых данных
2. Надежность, Smart-monitoring (временные ряды, построение регрессий и доверительных интервалов, NLP) – выявление отклонений в работе систем на основе прогнозных значений метрик. Кластеризация, локализация и выявление причин инцидентов, разбор жалоб, анализ логов
3. Автоматизация сервисных запросов внутреннего клиента (NLP, многомерная классификация) – Анализ сервисной заявки (комментарии пользователей и др.), маршрутизация на группы исполнителей, реком. системы, автоисполнение обращений
4. Управление мощностями, запасами, ресурсами (регрессионный анализ, генетические алгоритмы и др)
5. Управление архитектурой, качеством данных и справочной информацией (NLP, CNN и др.)
6. Создание чат-ботов для сервисов внутреннего клиента (NLP)

Collection-модели банков из топ-1

Предконтрактная стадия

1. Модели для оценки цессионных портфелей
2. Модели для workforce management
3. Модели для resource management

Верификация и обогащение данных

4. Модели для распознавания captcha при автоматизированном сборе данных
5. Модели для оценки вероятности найти информацию при помощи skip tracing
6. Модели для оценки вероятности принадлежности найденного контакта должнику
7. Модели для мэтчинга профилей физлица в различных онлайн-источниках

Робот-коллектор

8. Модели для определения наличия голоса в звуковом канале
9. Модели для распознавания речи
10. Модели для выбора направления движения по скрипту общения робота в зависимости от слов должника

Soft Collection

11. Модели для оценки мат.ожидания Recovery Rate
12. Модели для оценки вероятности дозвона по телефонному номеру
13. Модели для оценки ожидаемого финансового эффекта от звонка
14. Модели для прогноза длительности звонка
15. Модели для оценки вероятности получить обещание об оплате
16. Модели определения наличия автоответчика
17. Модели для определения результата переговоров с должником
18. Модели для рекомендации оптимальных мотиваторов
19. Разговорные модели для текстовых коммуникаций
20. Модели для динамического управления процессами взыскания на базе RNN

Hard / Legal / Executory Collection

21. Модели для оценки мат.ожидания Recovery Rate
22. Модели для оценки финансового эффекта от выезда сотрудника выездного взыскания

Примеры внедрений в банках

Post Mail Classifier

HOME CREDIT BANK

В банке Хоум Кредит построили AI, который позволяет разделять почтовую корреспонденцию по тематикам.

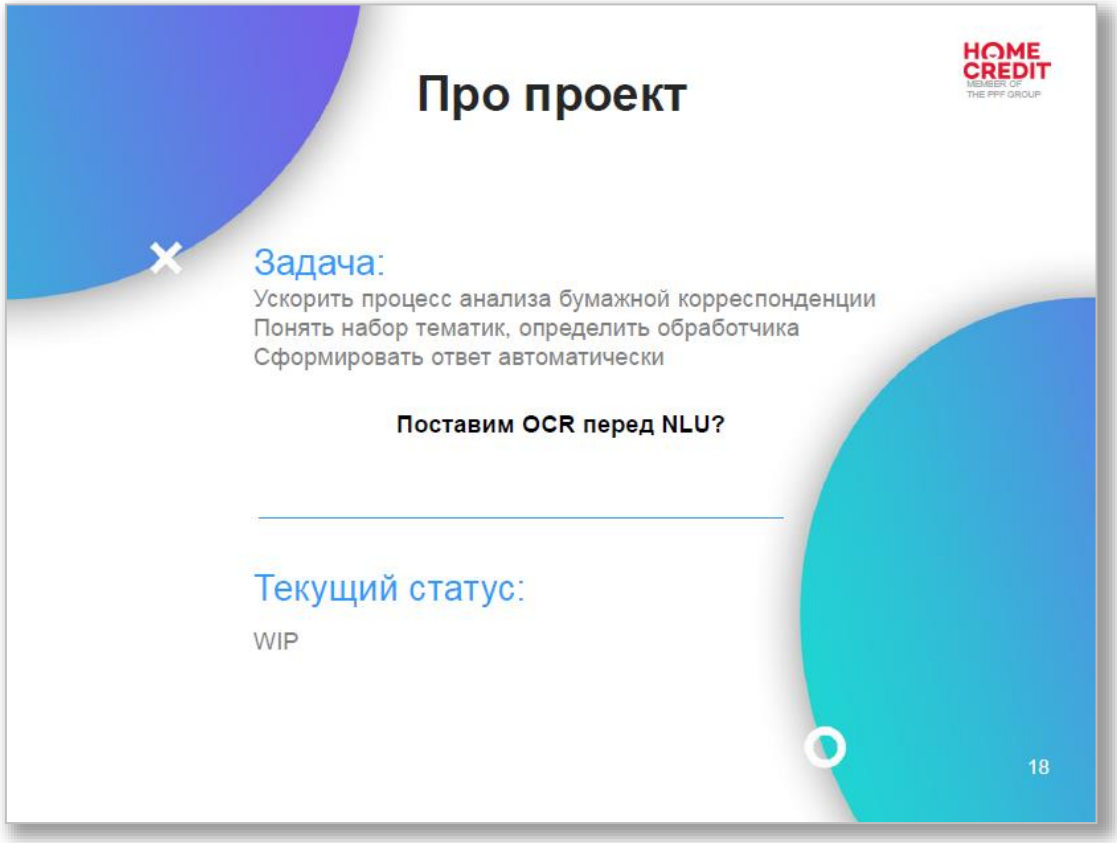
Технология позволяет распознавать печатные сканы писем, разделять их на 250+ тематик и формировать автоматический ответ на распознанные письма.

Для распознавания текста использовали Open Source технологию OCR: программу Tesseract от компаний Google.

End-to-end время: 10-40 сек

Качество распознавания печатных текстов: 78%

В основном тексты «юридического стиля»



2018. *Frankenstein Forum & Awards. The Retail Finance.*

Примеры внедрений в банках

Голосовая биометрия

Tinkoff Bank

В банке из топ-26 разработали свою собственную голосовую биометрию, которую внедрили на процессах Колл-центра и Коллекшн (идентификация клиентов).

Рабочая группа: ~10 человек
Срок разработки: ~1,5 года
Запуск: 2018 г.

Боль:
Потребовались дорогие доработки ИТ-инфраструктуры.

Примеры внедрений в банках

Чат-боты

В рамках конференции SDSJ-2017 Сбербанк провел соревнования по разработке чат-бота. Призеры получили предложения работать в лаборатории AI Сбербанка.

В банке Хоум Кредит разрабатывают чат-бота на основе AI, который расширяет возможности «кнопочного» чат-бота. Разработка находится на стадии пилотирования.

В банке Тинькофф разработали чат-бота, который выдает подсказки на вопросы клиентов. В диалоге пока используют живых операторов, которым чат-бот также помогает с подсказками для ответов.

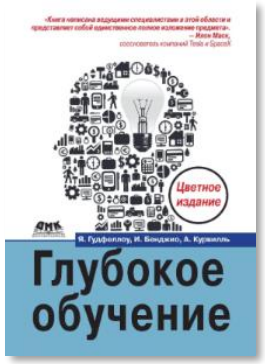
2017. *Sberbank Data Science Journey.*

Книги на русском языке

Глубокое обучение: погружение в мир нейронных сетей

Николенко С., Кадурин А., Архангельская Е.

СПб.: Питер, 2018. — 480 с.



Глубокое обучение

Ян Гудфеллоу, Иошуа Бенджио, Аарон Курвилль

ДМК-Пресс, 2018. — 652 с.

Библиотека Keras — инструмент глубокого обучения

Антонио Джулли, Суджит Пал

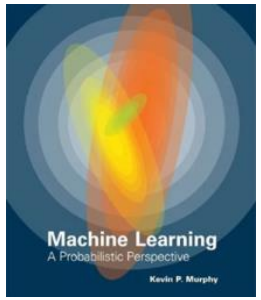
ДМК-Пресс, 2017. — 296 с.

Введение в машинное обучение с помощью Python

Мюллер А., Гидо С.

O'Reilly Media, 2017. — 392 с.

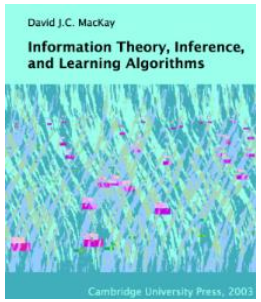
Книги на английском языке



Machine Learning: A Probabilistic Perspective

Murphy K.P.

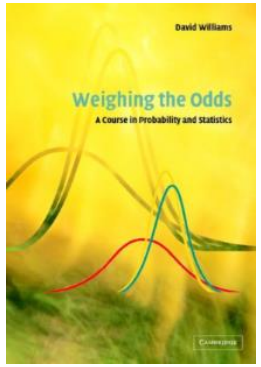
Massachusetts Institute of Technology, 2012. — 1067 p.



Information Theory, Inference, and Learning Algorithms

MacKay D.J.C.

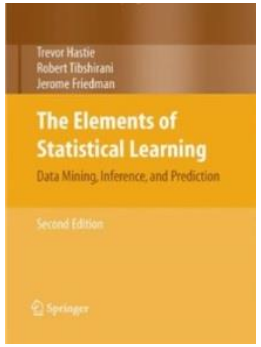
Cambridge University Press, 2003. — 640 p



Weighing the Odds: A Course in Probability and Statistics

David Williams

Cambridge University Press, 2001. — 568 p.



The Elements of Statistical Learning: Data Mining, Inference, and Prediction

Hastie T., Tibshirani R., Friedman J.

Second Edition (corrected 5th printing). — Springer, 2009. — 763 p.

Лекции и полезные ресурсы

https://www.youtube.com/playlist?list=PLJOzdkh8T5kp99tGTEFjH_b9zqEQiiBtC

Лекции ШАД от Константина Воронцова (д.ф.м.н. МФТИ, ВШЭ)

https://www.lektorium.tv/speaker/2691

Лекции Сергея Николенко (автор книги «Глубокое обучение: погружение в мир нейронных сетей»)

http://efimov-ml.com/

Сайт с лекциями Дмитрия Ефимова, а также с Python-скриптами (Jupiter Notebooks), презентациями и ссылками

http://deepbayes.ru/2017/

Лекции ВШЭ по Байесовским методам в глубинном обучении. Лекторы: Дмитрий Ветров, Дмитрий Кропотов, Евгений Соколов, Сергей Бартунов, Арсений Ашуха и др.

https://arxiv.org/

Ресурс с самыми свежими препринтами научных статей по ML/AI/DS (см. раздел Computer Science)

http://scikit-learn.org/
http://www.numpy.org/
https://pandas.pydata.org/
https://keras.io/
http://devdocs.io/tensorflow/

Подробные мануалы популярных библиотек в Python

https://stackoverflow.com/

Ресурс для обсуждения практических вопросов по программированию на Python и других языках

http://ods.ai/

Русскоязычное сообщество датасайнтистов