**Отзыв рецензента**

на выпускную квалификационную работу Благова Артема Анатольевича

**“Оценка надежности клиентов финансовой организации по их профилям в социальных сетях”**

В работе рассматривается проблема автоматической оценки заемщика на основе данных, которые предоставляет заемщик и той информации, которая находится в открытом доступе (социальные сети).

Описанная задача является одной из самых актуальных в кредитной сфере.

Большое количество людей, заинтересованных в кредитовании и высокая конкурентность рынка заставляет компании разрабатывать свои собственные, специфичные инструменты быстрой оценки заемщика.

Единого, коробочного решения не существует. Некий scoring ball (оценка заемщика в системе градации от 0 до 1000, где 0 - “самый плохой заемщик”, 1000 - “самый хороший заемщик”) предоставляет Национальное Бюро Кредитной Истории (НБКИ), но практика показала, что этот критерий не является качественным, однако является годным для “грубой” оценки.

Информационная безопасность и закрытость данных кредитных организаций также приводит к сложности создания единого решения. Кредитные организации неохотно делятся полной информацией о поведении заемщика, маржинальности, уровне дефолта и т.п.

Собрав информацию о заемщике, накопив собственную базу данных, решение задачи классификации заемщиков на “хороший” и “плохой” достижимо

благодаря современным инструментам IT и анализа данных - ставшие широко доступными средства machine learning (машинного обучения), построения независимых веб сервисов, интеграция с облачными, сторонними сервисами посредством API.

Как сказано в работе, ни одна из текущих, доступных систем оценки заемщика не “научилась” однозначно идентифицировать пользователя по голосу (если не было “первичного” контакта и факта соответствия голосового образца конкретному человеку). До сих пор, построенные модели не могут достоверно сказать, действительно ли человек, которому одобрен кредит, является тем, за кого он себя выдает, что является актуальной проблемой онлайн-кредитования.

Отдельной задачей стоит рассматривать использование в качестве дополнительных параметров модели, данные открытого доступа - социальные сети, сервисы учета штрафов, долгов, ФССП и т.д.

Методы анализа контента (фотографии, тексты постов, группы, сообщества, друзья и т.п.) известны достаточно давно, но до сих пор остается открытым вопрос их значения в характеристике человека (особенно с точки зрения узкого вопроса кредитоспособности человека).

Политика компаний, которые предоставляют такие данные не всегда позволяет использовать данные в промышленном масштабе. К примеру, то, что можно увидеть свободно через экран браузера не может быть получено автоматически при помощи API.

Это затрудняет легальный сбор данных для анализа (в работе приводятся ссылки на официальные источники).

Так или иначе, задача оптимум заключается в увеличении числа одобрения и отказов, полученных автоматическим путем. Таким образом, позволяя снизить нагрузку на ручной процесс обработки. “Спорные” заемщики (не однозначно “хорошие”, не однозначно “плохие”) могут быть рассмотрены специалистами в ручную.

В работе рассматривается модель существующего и работающего решения, построенная с учетом современных инструментов приложения информационных и статистических наук:

*микросервисная архитектура, паттерн “Event bus”, инструменты машинного обучения (методы оптимизации обучения модели и т.д.).*

Данные для анализа поступают из нескольких источников, включая анкету заемщика, профиль социальной сети “ВКонтакте”. Интеграция с социальной сетью “ВКонтакте” происходит по API. Приведен список полей, доступных для анализа с их описанием.

Технически грамотно описан процесс анализа данных, получение работающей модели:

выбор данных, нормализация данных, “чистка” данных от корреляций, ранжирование данных по степени важности, построение модели инструментами машинного обучения, верификация модели. Недостаточно подробно описан процесс выбора классификатора для решения задачи.

Для имплементации модели выбран микросервервисный архитектурный паттерн, который является хорошим решением для инкапсулирования логики, повышение отказоустойчивости и работы с большими нагрузками. Микросервисный паттерн является современным, технологичным решением, использующим новый протокол gRPC. Подробно описана архитектурная модель решения: общая архитектура, бизнес-логика микросервиса, модель “выходных” сообщений сервиса. Приводятся ссылки на книги, статьи с подробных описанием используемых инструментов.

Выводы работы достаточно однозначны, корректны и достаточно пояснены.

Тем не менее, я бы хотел отметить, что в данной работе рассматривается лишь один пример анализа данных из соц. сетей. С другой стороны, глубокий анализ контента затрагивает большое количество смежных прикладных наук, как data mining, text mining, picture recognition и т.д. и таким образом не может быть покрыто одной работой и знаниями одного специалиста.

**К положительным сторонам работы отношу:**

1. актуальность проблемы
2. использование современных, технологичных, сложных инструментов
3. грамотный технологический процесс
4. хорошо продумана и описана архитектура решения
5. большое количество ссылок на сторонние ресурсы, литературу
6. работа написана простым, доступным языком с иллюстрированием текста, где необходимо

**К отрицательным:**

1. Тема не раскрыта полностью (описана одна соцсеть с достаточно “скудным” набором данных)
2. Отсутствие примеров исходного кода (особенно при обучения модели)

Несмотря на имеющиеся замечания, считаю, что работа удовлетворяет требованиям к ВКР бакалавра и заслуживает положительной оценки.

**Рецензент:**

к.ф-м.н., м.н.с.

ФТИ им. Иоффе РАН,

Мишуров Игорь Юрьевич

руководитель проектного офиса компаний “Девим”,

руководитель направления автоматические скоринговые системы.

28.05.2018