**Clients risk category prediction using diversified credit scoring model**

**Chioroglo Alexandr 1\*,**

*1 Technical University of Moldova, FCIM. ISA departament, group* *IA-213, Chișinău, Republic of Moldova*

\*Author: Chioroglo Alexandr, alexandr.chioroglo@iis.utm.md

***Abstract:*** *Properly analyzing the financial portray of a client is a substantial part of building a profitable banking institution. Nowadays, such an important inspection may be near fully automated using credit scoring algorithms based on machine learning techniques and models. These algorithms are built relying on information from credit bureaus, businesses own settings and biases, government revenue services data etc. This topic describes an example of building that kind of machine learning model with corresponding proofs, hypothesis analysis,*

***Keywords:*** *Credit scoring, credit history analysis, risk category, price category, diversified scoring,*

**§1 Introduction**

Кредитный скоринг представляет собой статистический анализ, который проводится

кредитующими организациями чтобы определить кредитоспособность отдельно взятого человека или небольшого бизнеса. Кредитный скоринг автоматизирует принятие решений по выдаче или невыдаче кредита, может также определять определённые условия для займа.

Кредитный скоринг основывается на различных группах данных, которые берутся из различных легальных источников: налоговых служб, которые предоставляют данные по специальной лицензии, кредитных бюро, данных, указанных человеком в декларации. Данные, используемые в скоринге включают в себя:

1. История платежей по кредитам (суммы, просрочки, сроки, даты, предыдущие запросы в кредитное бюро), помогает анализировать кредитное поведение человека, склонность к выплате/невыплате платежей по кредитам.
2. Текущие задолженности (суммы, типы кредитов, сроки, степень задолженности, процентные ставки) – помогает анализировать финансовую грамотность заёмщика, степень его кредитной нагрузки, анализировать процентные ставки, предложенные другими банками.
3. Информация об источниках дохода субъекта – сфера деятельности, заработная плата, иные источники дохода, информация об инвестициях и т.д

Данный список не является полным и отражает лишь основные данные, используемые для принятия решений в оценке рисков по займам. Эти данные выступают основными в таких моделях скоринга как FICO© и Vantage Score ©. Эти модели являются диверсифицированными, что объединяют в себе все типы скоринга: Application Scoring, Behavioral Scoring, Fraud Scoring, Collection Scoring.

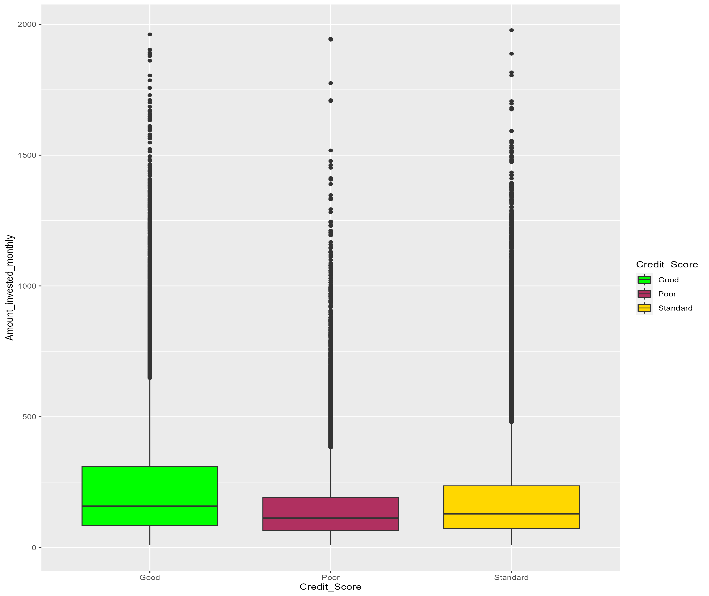
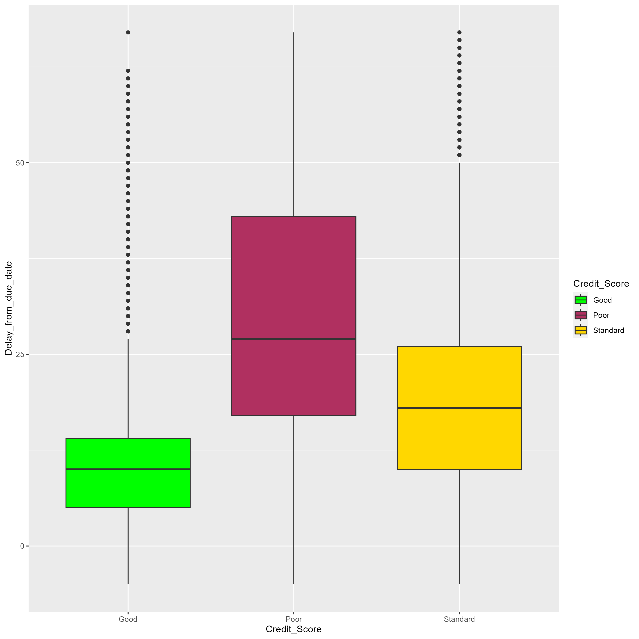
Целью данной статьи является e2e описание процесса построения модели машинного обучения, которая может быть использована на предприятии для оценки рисков при выдаче кредитов. В качестве данных для модели были взяты данные о платеже одной из финансовых компаний, деятельность которой близка к деятельности кредитного бюро. В процессе работы с помошью различных инструментов будет совершен переход от сырых данных к модели, которая способна принимать на вход те же данные, но предсказывая категорию риска клиента по диверсифицированноймодели

**§2 Materials and methods**

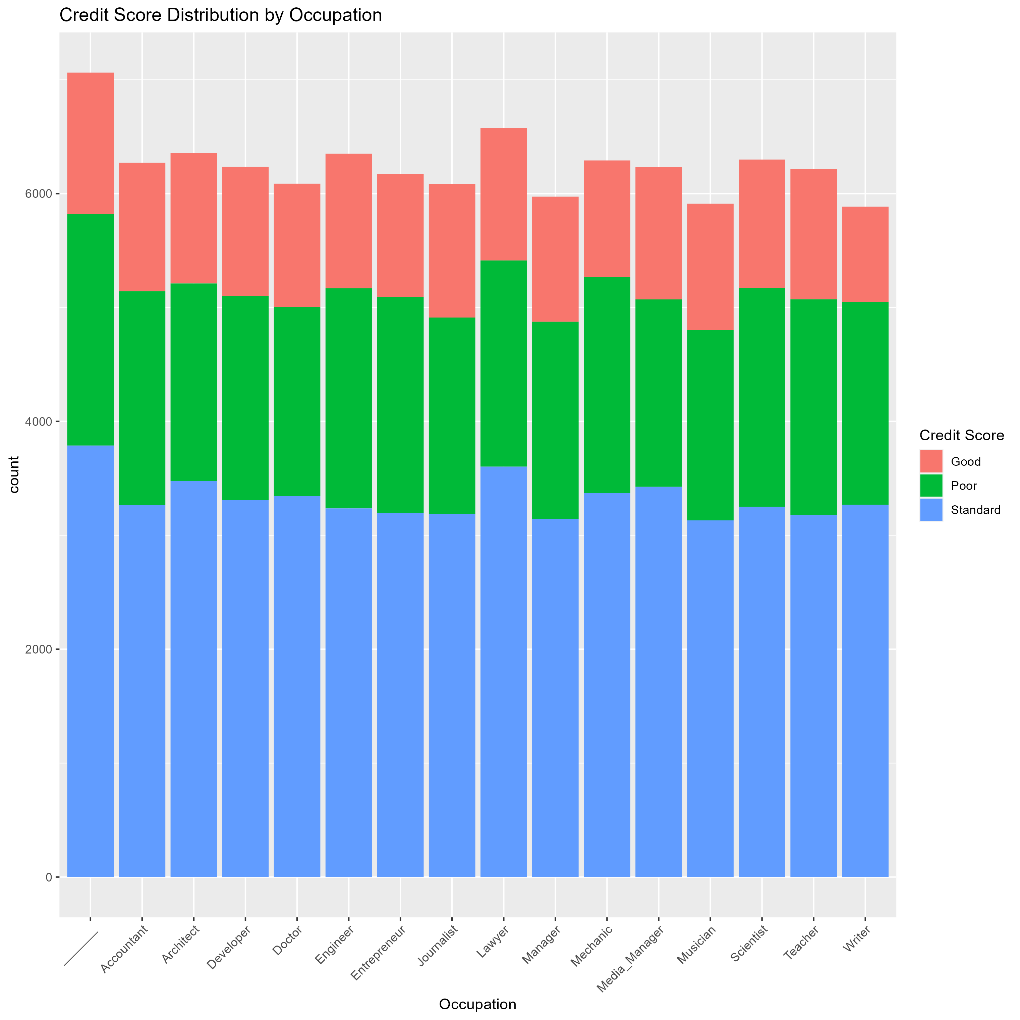
Для работы над моделью был взят учебный набор данных с сайта Kaggle© (https://www.kaggle.com/datasets/parisrohan/credit-score-classification), который представляет собой информацию о клиентах компании, преимущественно связанную с их кредитной активностью. Информация делится на три категории:

1. Персональные данные – (возраст, имя, SSN, идентификаторы, профессия etc).
2. Финансовые данные – (инвестиции, зарплата,годовой/месячный доход).
3. Кредитная история – (количество займов, остаточные долги, процентная ставка, количество запросов в кредитное бюро, кредитная нагрузка)

Для проведения анализа данных был использован язык программирования R, который предоставляет широкий спектр инструметов стандартной библиотеки. Различные модули позволяют расширить функционал. В основном, дополнительные модули устанавливались для того, чтобы провести нестандартную визуализацию данных (highcharter, plotly). Для создания модели машинного обучения был использован язык программирования Python с библиотеками scikit-learn, tensorflow для обучения модели и matplotlib/seaborn для визуализации данных. После очистки данных, были очищены колонки, которые не оказывают влияние на финальную переменную. Переменной, которая не оказывает влияние на финальный результат может считаться та переменная, что не отличается по основным статистическим показательям (summary), графически с помощью boxplot (ящик с усами).



***Figure 1.*** *Пример двух переменных, анализируемых boxplot-ом. Слева – важная пременная, т.к категория скоринга прямо зависит от максимальной просрочки по оплате. Явно просматриваются отдельные кластеры. Справа же наблюдается не сильная кластеризация, все переменные примерно в одной и той же области.*



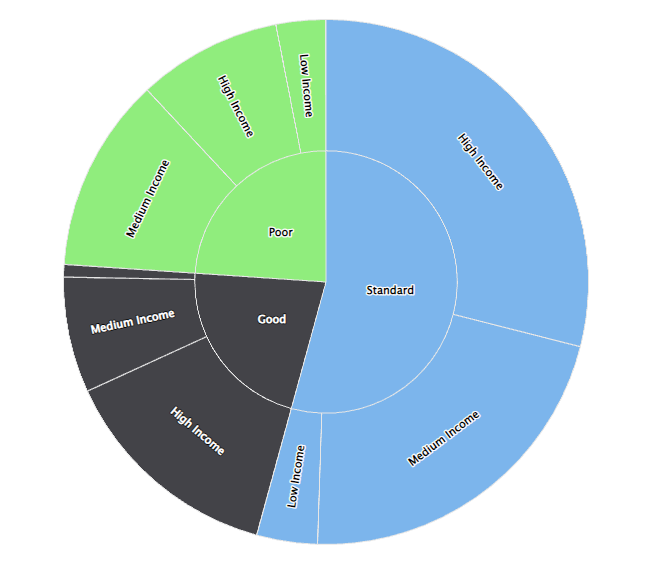
***Figure 2.*** *На данном графике изображено распределение категории скоринга внутри групп клиентов с определённой профессий. Согласно графику видно, что внутри кластеров профессий примерно одинаковое распределение внутри категорий. Это может быть подтверждено с помощью критерия Пирсона (Chi-Squared measure).*

**§3 Results**

В первую очередь был произведен анализ и приведение категориальных переменных к численным для того, чтобы модели было удобнее разделять клиентов на различные категории как признаки. Также были нормализованы данные и убраны выбросы, Выбросы были идентифицированы и убраны или заменены по официальному определению, с использованием интерквартильного расстояния. Вкратце, был взят первый и третий квартиль, было посчитано расстояние в полтора межвартильных расстояния. Все значения, что выходят за предел больший этого значения от медианы – являются выбросами. Некоторые категориальные переменные были разделены с помощью стандартных методик.

В результате предварительного анализа данных были выведение определённые гипотезы, которые были опровергнуты или приняты на основании определённых исследований, подкреплённых графически и статистически. В реальных условиях, тем не менее, необходима консультация со специалистами в предметной области.

* Зарплата не оказывает влияние на финальный результат
* Профессия не оказывает влияние на финальный результат
* Месячные инвестиции человека оказывают влияние на финальный результат



**Figure 3.** Распределение категорий дохода внутри категорий кредитного скоринга.

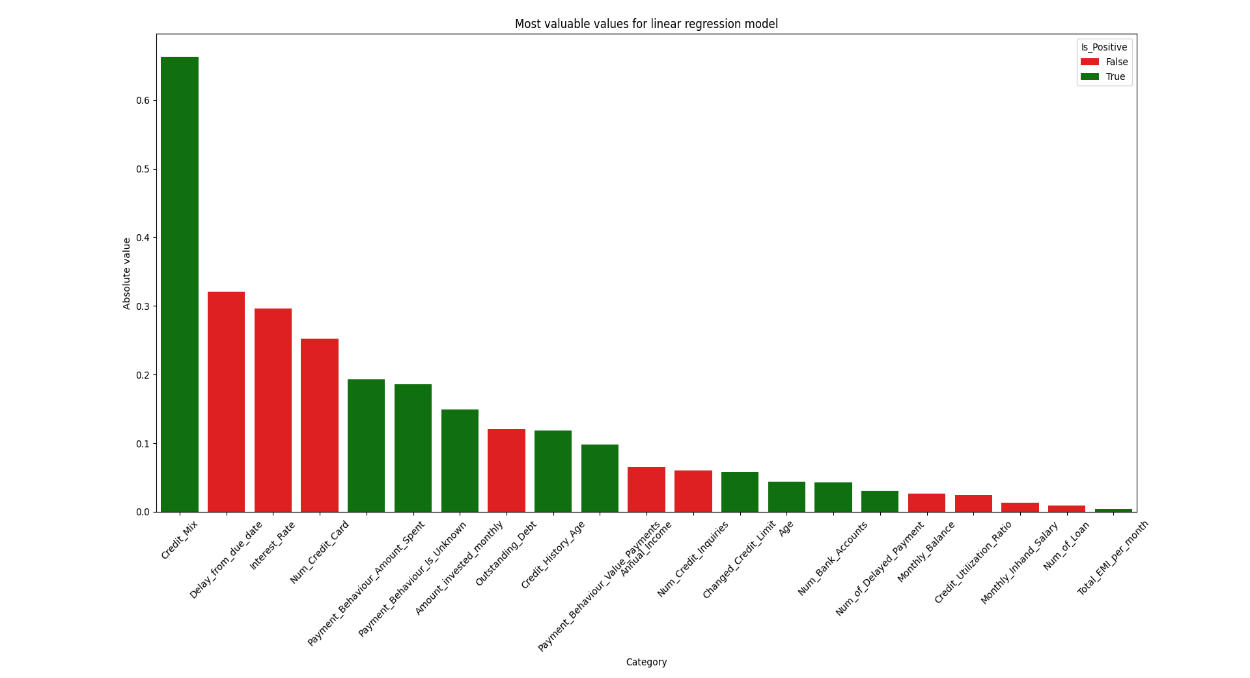
На данном графике можно увидеть распределение клиентов по категориям зарплаты внутри категорий кредитного скоринга. Видно, что в каждом кластере данных примерно одинаковая доля клиентов с определённым уровнем дохода, что свидетельствует о потенциальной независимости результата от дохода. Но, учитывая анализ предметной области, опираясь на статьи, а также на законодательство – мы не можем упускать из виду эту переменную. Значение дохода должно удовлетворять определённым минимальным требованиям для выдачи займа по закону (лимит на 40% или 55% от дохода в месяц в угоду оплаты ежемесячных задолженностей по погашению кредитов). Ввиду этого, было принято решение оставить данную переменную в расчётах.

В случае с данной задаче, в качестве моделей машинного обучения стоит выбор – нейронная сеть или же логистическая регрессия.

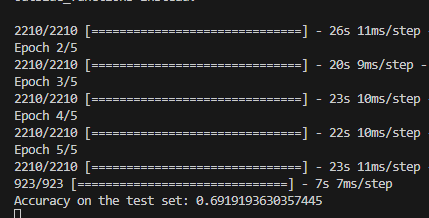
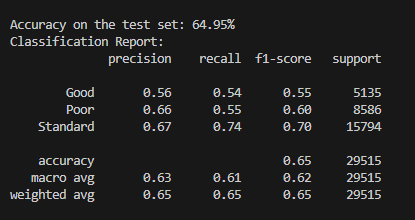
**Нейронная сеть** способна запоминать различные нелинейные паттерны и взаимосвязи, если использовать правильные алгоритмы (LSTM), что может быть трудным для традиционных линейных моделей, но, они запрашивают довольно большого количества данных на вход. Но, мы можем ограничиться одним алгоритмом.

**Логистическая регрессия** в единственном экземпляре хорошо справится с задачей бинарной классификации, соответственно необходимо составить три модели, каждый раз запускать три алгоритма и проверять вероятность принадлежности к определённой категории риска, брать наибольшую из них и предоставлять как ответ на вопрос, поставленный задачей.

Логистическая регрессия, к тому же, является довольно интерпретируемой на человеческий язык, что упрощает задачу объяснения выводов, сделанных алгоритмом менеджменту. Банки - довольно консервативные организации, которые предпочитают опираться на цифры, формулы. Логистическая регрессия имеет чёткую формулу, нейронная же сеть представляет собой огромный blackbox , решения которой зачастую трудно объяснить заказчику.



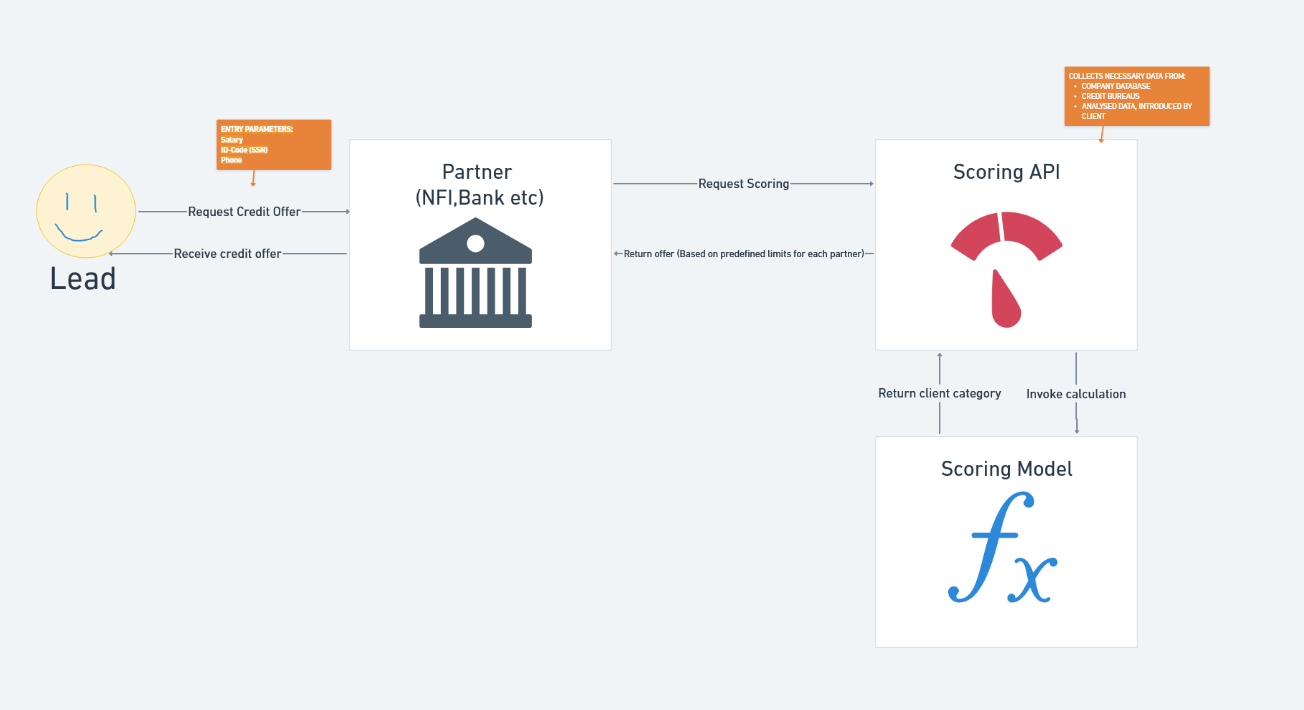
**Figure 4.** Визуализация самых значимых переменных в модели линейной регрессии. По оси Y – значение коэффициента при входых данных, красным или зелёным цветом обозначено положительно ли влияет на категорию риска фактор или отрицательно.

**Figure 5.** Метрики согласно моделей. Слева – нейронная сеть, справа – модель линейной регрессии.

Модель может быть потенциально улучшена с помощью техник кросс-валидации, дебаггинга. Выбор грамотных пресетов при обучении (learning rate, biases etc), смена алгоритмов оптимизации может также помочь в достижении большей точности, но необходимо быть аккуратным и проверять модель на предмет переобучения с помощью различных метрик (R-square,Chi-Squared measure, как пример).

Имея на руках модель, которая обладает точностью в 70%, можно думать об архитектуре и сфере применения данной модели. Данная модель может быть использована как алгоритм для принятия решений на предприятии. Ниже приведена пилотная схема интеграции алгоритма в работу банка.



**Figure 6.** Схема интеграции алгоритма в банковскую экосистему. Банк (конкретный или партнерский) будет подключаться к модели посредством API, который будет заниматься фильтрацией и валидацией входных данных, после чего отправлять запрос скоринговой модели, результат которой будет логироваться для последующего анализа со стороны бизнеса и возможного использования как датасета в будущем.

**§4. Discussion**

Скоринг — это система оценки заемщика, с помощью которой банки и крупные микрофинансовые организации (МФО) могут предсказать, насколько аккуратно человек будет выплачивать кредит. В основе прогноза лежат математические расчеты и статистика.

Считается, что люди с похожими привычками примерно одинаково обращаются с финансами. Поэтому банки и МФО сравнивают потенциальных заемщиков с другими людьми, которые уже брали похожие кредиты. Они строят скоринговые модели, в которых платежеспособность человека оценивает компьютерный алгоритм. Эта работа описывает схему, по которой банки строят свои кредитные модели. Схематика построения процессов основана на опыте работы автора в микрофинансовых организациях, консультациях со специалистами, а также чтением литературы и статей. Минусами этого исследования являются малый акцент на алгоритмах машинного обучения, в процессе работы могло быть уделено больше внимания различным моделям (K-Nearest-Neighbours,Decision Trees). Максимально достигнутая точность – 70%, что могло бы быть улучшено, так как при случайном переборе точность является 33%, разница в 40% между случайным выбором навряд-ли убедит банк пользоваться данным алгоритмом, точность и остальные метрики должны чётко указывать на близкую к 80-90% точности. Сильной стороной данной работы является профессиональная работа с предобработкой набора данных, были использованы лучшие практики для обработки – обработаны все категориальные переменные (упорядочиваемые и неупорядочиваемые) и приведены к однообразному виду. Были использованы инструменты для анализа данных, что используются в крупных компаниях (R, python etc), что может довольно убедительно повлиять на авторитет исследователя. В случае же с реальным банком, дата сайнтисту будет предоставлена свобода собирать данные, которые ему угодны. Это же исследование во многом полагалось на уже наличные данные, в реальной компании дата сайнтист имел бы больше свободы чем при учебной задаче.

**§6. Conclusion**

Целью данного исследования было приведение примера процессов создания модели кредитного скоринга в банковских организациях, были показаны четыре главных этапа:

1. Предобработка собранных данных
2. Составление гипотез и их опровережение, принятие.
3. Генерация архитектуры модели, оценки точности.
4. Интеграция модели в приложение

Эта статья может помочь дата сайнтистам в расширении кругозора в финансовой сфере как пример поставленных процессов в различных компаниях для того, чтобы лучше быть готовым к тому, с чем им предстоит столкнуться на реальном месте работы.

**§7. Useful links:**

* **DATASET:** Credit Score Classification Competition // KAGGLE // <https://www.kaggle.com/datasets/parisrohan/credit-score-classification>
* **SOURCE CODE:** CreditScoringStudyingUtm repository // Chioroglo // <https://github.com/chioroglo/credit-scoring-studying-utm>

**LITERATURE SUPPLEMENT:**

1. <https://www.investopedia.com/terms/c/credit_scoring.asp> - Credit Scoring // Thomas Brock, The Investopedia Team © // March 2023
2. <https://www.forbes.com/advisor/credit-score/fico-vs-vantagescore-credit-scores-whats-the-difference/> - FICO vs VantageScore Credit Score : What’s the difference? // Michelle Black, Dia Adams, Forbes Media LLC © // March 2021
3. <https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html> - How to deal with categorical data for Machine Learning // Shelvi Garg, Data Scientist at Spinny // August 2022
4. <https://builtin.com/data-science/disadvantages-neural-networks> - 4 disadvantages of Neural Networks // Niklas Donges, Built In (c) 2023 // August 2023
5. <https://fincult.info/article/skoring-kak-banki-i-mfo-reshayut-davat-li-vam-kredit/> - Скоринг: как банки и МФО решают, давать ли вам кредит // ФинКульт, проект Национального Банка Российской Федерации // October 2022
6. *legis.md: Registrul de stat al actelor juridice. (RU)*. Закон №202 от 12-07-2013 о кредитных договорах с потребителями. Редакция от 30.06.2023. Вступает в силу с 03.07.2023 © Ministerul Justiției, 2023. [цитировано 12/08/2023]. Доступно: https://www.legis.md/cautare/getResults?doc\_id=137812&lang=ru
7. *legis.md: Registrul de stat al actelor juridice. (RU)*. Постановление №20/5 от 20-05-2022. Редакция 20-05-2022 Об утверждении Регламента о требованиях ответственного кредитования, применяемых к небанковским организациям. © Ministerul Justiției, 2023. [цитировано 12/08/2023]. Доступно: <https://www.legis.md/cautare/getResults?doc_id=131857&lang=ru>
8. *Organizații de creditare nebancară –* Comisia Națională a pieței financiare Republica Moldova © 2023 – [цитировано 9/3/2023] - <https://www.cnpf.md/ro/organizatii-de-creditare-nebancara-6456.html>
9. *What is an Application Score,* ***-***© Credit Reporting Agency Ltd 1999-2023. All rights reserved*. –* [цитировано 30/11/2023] - <https://www.checkmyfile.com/jargon/application-score.htm>
10. *temabiz.com -* “Эффективная процентная ставка по кредиту”. © 2013-2020, Бизнес идеи, бизнес планы, школа бизнеса. [цитировано 12/08/2023]. Доступно: <https://temabiz.com/finterminy/effektivnaja-procentnaja-stavka.html>