

**Приложен изкуствен интелект**

**Тема на проекта: Long-Short Term Memory (LSTM) for credit risk assessment**

***Изготвили: Ростислав Банков и Александър Георгиев,***

***ФН: 471221117 и 471221116***

***Преподавател: Доц. д-р инж. Александър Фремов***

***Съдържание:***

**1. Въведение**…………………………………………………………………………………………………………….2

* 1. Идея………………………………………………………………………………………………………….2
  2. Терминология..………………………………………………………………………………………..2
  3. Подходи и Методи…………………………………………………………………………………..2

**2. Теория**……………………………………………………………………………………………………………………3

2.1 Описание на алгоритмите……………………………………………………………………….3

2.2 Реализация на основни инструменти за работа с данни……………………….6

**3. Примери**………………………………………………………………………………………………………………..7

3.1 Описание на задачата………………………………………………………………………………7

3.2 Описание на данните……………………………………………………………………………….8

3.3 Обработка на данните……………………………………………………………………………..8

3.4 Моделиране……………………………………………………………………………………………..8

**4. Експеримент**…………………………………………………………………………………………………………..9

4.1 Резултати……………………………………………………………………………………………………9

**5. Заключение**……………………………………………………………………………………………………………11

**Въведение**

**Идея:**

Кредитната рискова оценка е от съществено значение за финансовите институции и кредитодателите, които желаят да предвидят вероятността от невъзстановение на заеми от страна на клиентите. С течение на времето, се развиват различни методи и подходи за подобряване на точността на тези прогнози. Един от подходите, който привлича вниманието, е използването на реактивни невронни мрежи, като например Long-Short Term Memory (LSTM).

**Терминология:**

LSTM (Long-Short Term Memory): Това е вид рекурентна невронна мрежа с архитектура, предназначена за решаване проблема с изчезващия или прекалено увеличаващият се брой данни при обучение на RNNs.

Кредитен Риск: Вероятността от загуба, свързана с невъзможността за връщане на заемените средства от страна на заемополучателя.

Подходи & Методи: Моделът използва данни за доход, кредитна история, отношение дълг/доход и възраст, за да предскаже кредитния риск. В този случай и корелационната матрица също се взема предвид като допълнителен признак.

**Подходи & Методи:**

Целта на нашия код е да предостави примерeн LSTM модел и неговото обучаване, използвайки различни библиотеки като Torch, Pandas и NumPy. Данните за доход, кредитна история и др. се подготвят и map-ват в числови стойности. Данните от примерния файл се преобразуват на PyTorch tensor-и, след което се изчислява корелационна матрица, като се отчита влиянието на различните фактори.

Обучение и Оценка: Моделът се обучава чрез оптимизация на функцията за грешка (CrossEntropyLoss) и се оценява върху тестови данни. Обучението се извършва чрез метода на задния ход (backpropagation).

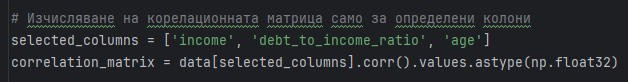
Предсказания и Резултати: Моделът предсказва категорията на кредитния риск за всяка индивидуална заявка. Резултатите се интерпретират и анализират за вземане на интелигентни решения за заемане.

**Теория**

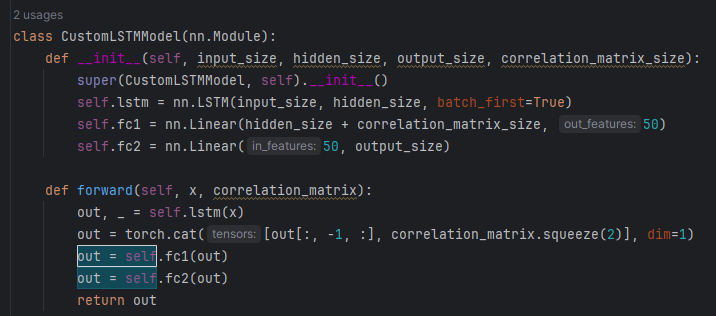
**Описание на алгоритмите**

В програмата има различни видове алгоритми и функции, които в контекста на кредитната оценка на риск сме базирали около LSTM модела като основа на нашия код. Тези рекурентни невронни мрежи (RNN), които са проектирани да работят с последователни или времеви примерни данни, които биват прочетени от програмата. Ето я и последователността на работата на програмата спрямо функциите и алгоритмите в нея:

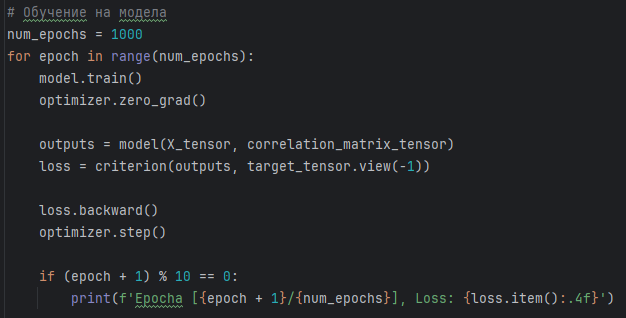
**1**.**Изчисляване на корелационната матрица**

Алгоритъмът включва изчисляване на корелационна матрица, която представлява мярка за връзката между различните колони от данните. Тази матрица се използва по-късно като допълнителен признак в модела.

**2. Създаване на LSTM модел**

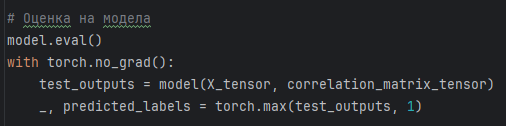
Дефинира се и инициализира LSTM моделът (CustomLSTMModel), който включва LSTM слой, отговорен за запазване на контекста от предходните стъпки.

**3. Обучение на модела**

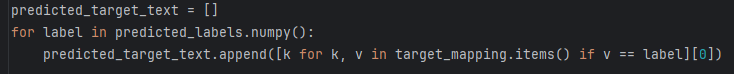
Обучението на модела се извършва чрез подаване на входните данни и корелационната матрица на LSTM модела, следвано от грешка и оптимизация чрез крос-ентропия и Adam оптимизатор.

**4. Оценка на модела**

След успешното обучение, моделът се оценява върху същите данни, за да направи предсказания относно кредитния риск.



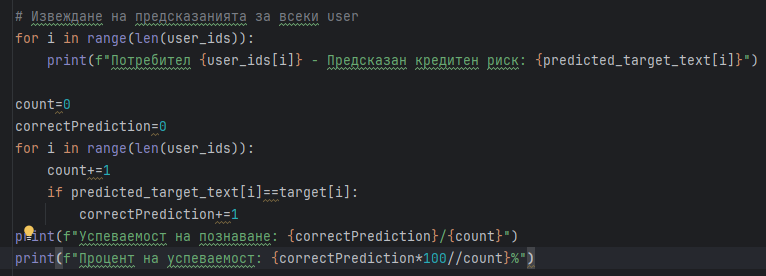
**5. Преобразуване на предсказанията**

Получените числови предсказания се конвертират обратно към текстуални стойности, които представляват различните нива на кредитен риск.

**6. Извеждане на предсказанията за всеки user и изчисляване на успеваемостта**

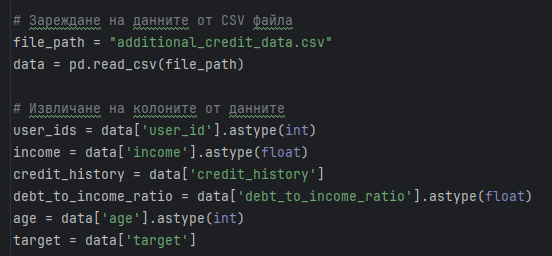
За всякакъв потребител, моделът извежда предсказанията за нивото на кредитен риск. Това включва изчисляване на успешността на модела, като се сравняват предсказанията със реалните стойности и се изчислява процентът на правилни

предсказания.



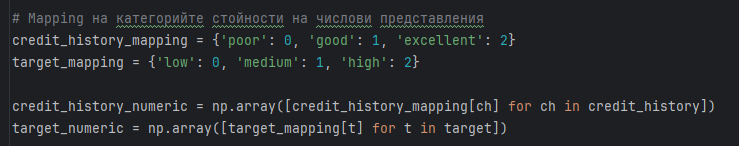
**Реализация на основни инструменти за работа с данни**

1. **Зареждане и извличане на данните:**

Алгоритъмът започва със зареждане на финансовите данни от CSV файл. Този етап включва използването на библиотеката pandas, която предоставя удобни инструменти за манипулация на данни и анализ. След това те биват извлечени.

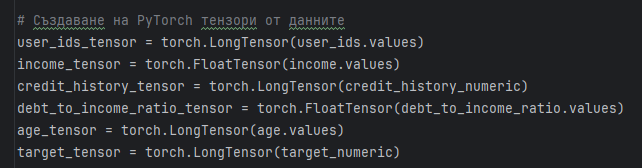
1. **Mapping на някои категории от стойности:**

Следващата стъпка включва преобразуването на категорийни стойности като 'credit\_history' и 'target' в числови представления. Този процес се извършва чрез мапване, където всяка категория се превръща в уникален числов идентификатор



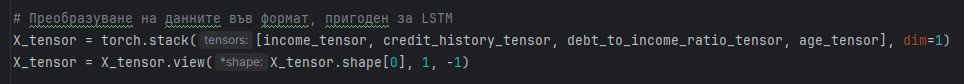
1. **Създаване на PyTorch тензори:**

Данните се преобразуват в PyTorch тензори, които са основната структура за представяне на данни в PyTorch. Всяка колона от данните, като 'income', 'debt\_to\_income\_ratio', и 'age', се превръща в съответен тензор.



1. **Преобразуване на данните във формат, пригоден за LSTM:**

Данните се структурират във формат, съвместим с LSTM, като се стекират в триизмерен тензор. Този формат е подходящ за последователност от данни, както и за запазване на връзката между стъпките.



**Примери**

**Описание на задачата**

Целта на задачата е да се разработи модел, който способен да предсказва нивото на кредитен риск за даден потребител, базирано на различни финансови характеристики. Този модел може да бъде полезен за финансови институции при вземането на решения относно одобряване или отказ на кредити.

**Описание на данните**

Входните данни представляват таблица с различни характеристики на потребителите, включително доход, кредитна история, дължим доход, възраст и текущ кредитен риск (целева променлива).

**Примерна таблица:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| user\_id | income | credit\_history | debt\_to\_income\_ratio | age | target |
| 1 | 50000 | good | 0.2 | 35 | low |
| 2 | 60000 | excellent | 0.3 | 40 | medium |
| 3 | 75000 | good | 0.25 | 30 | medium |
| 4 | 80000 | poor | 0.4 | 28 | high |
| 5 | 55000 | excellent | 0.35 | 32 | low |
| 6 | 70000 | good | 0.18 | 37 | medium |
| 7 | 48000 | poor | 0.38 | 41 | high |
| 8 | 65000 | good | 0.22 | 36 | medium |
| 9 | 72000 | excellent | 0.28 | 29 | medium |
| 10 | 58000 | good | 0.33 | 43 | low |

**Обработка на данните**

Данните се обработват чрез mapping на категории стойности в числови представяния, създаване на PyTorch tensors и изчисляване на корелационната матрица. Тези стъпки са от съществено значение, за да се гарантира, че моделът е в състояние да обработва и извлича полезна информация от данните.

**Моделиране**

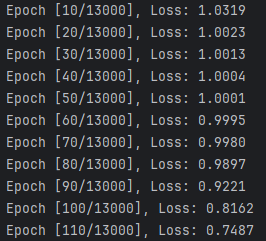
Използвайки библиотеката PyTorch, се създава и обучава LSTM модел, който може да улавя зависимостите във времевите данни. Моделът се конфигурира от 50 неврона и два за класификация. Този пример подчертава важността на обработката на данни и моделирането с LSTM за успешното решаване на задачи, свързани с кредитната оценка на риск.

**Eксперименти**

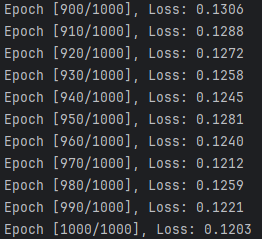
**Резултати**

От долуспоменатите изображения на резултатите можем да забележим как успеваемостта на LSTM модела варира спрямо броя на направените епохи и как постепенно той се подобрява:

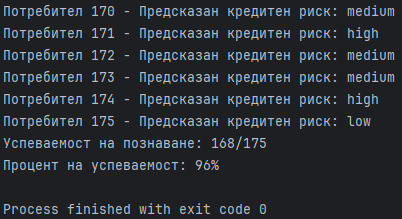
* При 100 преминати епохи виждаме, че грешките при предсказаните кредитни рискове са относително висока



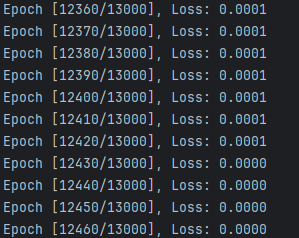
* При вече 1000 преминали епохи забелязваме значително намаляване на грешките



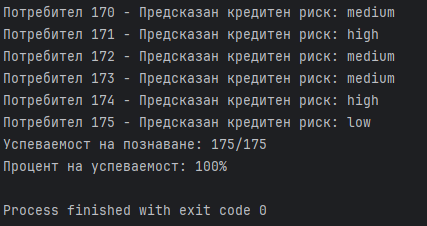
* Грешката не е напълно елиминирана и затова накрая в резултати виждаме непълна успеваемост на предсказаните кредитни рискове



* В следващия пример ще можем да забележим, че при значителен брой повече епохи (в този случай 13000) LSTM моделът успява да сведе грешките до 0 след четвъртия знак от запетаята



* В такъв случай можем да се насладим на 100% успеваемост при предсказване на кредитен риск



**Заключение**

Представеният проект, фокусиран върху оценката на кредитен риск с използването на Long-Short Term Memory (LSTM) модел, представлява иновативен и цялостен подход в областта на финансовия анализ и машинното обучение. Разработката на проекта обхваща внедряването на идеята, изграждането на модела и прилагането на специфични техники за обработка на данни и обучение.

Процесът на изграждане на LSTM модела е осъществен със структуриран и систематичен подход. Методите за мапиране и преобразуване на данни, използвани във въвеждащата фаза, позволяват успешната адаптация на финансовата терминология към машинното обучение. Съчетаването на тези две области създава уникален инструмент за анализ на данни във финансовия сектор.

Реализацията на LSTM модела, който интегрира корелационна матрица и статистически параметри, предоставя значително по-широкообхватен анализ в сравнение с традиционните модели. Този подход подчертава значението на включването на различни финансови показатели за по-добро предсказване на кредитния риск.

Използването на Cross-Entropy Loss и оптимизатора Adam подобряват ефективността на обучението. Процесът на оценка на модела включва анализ на предварително определените маркери за успешно предсказване.

Резултатите от тестовете демонстрират висока степен на успешно предсказване на кредитния риск, което подчертава потенциала на предложения модел. Точността на модела се измерва във възможния процент успех, допринасяйки към по-широкото разбиране на рисковете във финансовата среда.