Проект. ИИ и FinTech: прогноз лучшего финансового предложения.

1. Введение.

В настоящее время методы математического моделирования активно применяются в финансовом секторе. Такие учреждения, как банки, страховые и микрофинансовые организации в своей работе уже сейчас используют эффективные математические подходы, среди которых выделяются методы искусственного интеллекта и статистического моделирования. Самыми яркими решениями, которые базируются на упомянутых математических методах, являются противодействие финансовому мошенничеству, обработка финансовых документов, обслуживание банкоматов, использование голосовых помощников.

2. Формулировка проблемы.

Хотя упомянутые и многие другие технологические решения широко используются в финансовой сфере, но всё ещё существует много вопросов в одном из самых важных финансовых направлений, а именно, в оценке платёжеспособности заёмщика. Каждый день многие россияне рассматривают кредитные предложения банковских организаций и подают заявки на одобрение. Но по итогу после долгого времени рассмотрения получают отказ. Кроме того, сам процесс рассмотрения заёмщика может потребовать силы немалого числа специалистов. Поэтому в данном проекте мы разрабатывали методы, которые смогли бы облегчить принятие решений об одобрении либо отказе займов по конкретным транзакциям заёмщиков.

В ходе работы нашей команде были высланы два набора данных о транзакциях заёмщиков в микрофинансовые организации. В первом наборе каждая транзакция какого-то конкретного заёмщика представлялась вектором из нескольких десятков характеристик, среди которых, например, были: названия региона и города заёмщика, оффер — название организации, в которую отравлена транзакция. Во втором наборе транзакции заёмщиков содержали в себе характеристики из первого набора и дополнительно ответы заёмщика на вопросы анкеты: сумма займа, срок займа, способ получения займа, доход заёмщика, наличие просрочек у заёмщика. На эти вопросы заёмщик отвечал во время составления своего обращения. Поэтому план действий нашей команды состоял в следующем. Вначале мы анализировали первый набор данных, находили в нём информативные и неинформативные характеристики, создавали новые характеристики, которые потом могли бы применить другого набора данных. Далее мы анализировали второй набор данных, занимались конструированием новых характеристик, по которым можно было бы прогнозировать одобрения либо отказ займа для конкретной транзакции. Для этого шага мы использовали результаты анализов двух высланных набор данных.

3. Анализ первого набора данных.

Первый набор данных содержал в себе более 110 тысяч транзакций, каждая транзакция содержала в себе 21 характеристику. На рис. 1 изображена для примера часть данных.

Оффер	Время конверсии	Статус	Поисковая система	Плохая кредитная история	Возрастная группа	Безработные	Читают отзывы перед оформлением	Город	Тип устройства	UserAgent
Быстроденьги	16.07.2022, 22:14	Одобрено	Яндекс	NaN	NaN	NaN	NaN	Kimovsk	Планшет	Mozilla/5.0 (Linux; arm_64; Android 11; RMX326...
Кэш-Ю Финанс	16.07.2022, 16:23	Одобрено	Яндекс	NaN	NaN	NaN	NaN	Tula	Мобильный телефон	Mozilla/5.0 (Linux; arm; Android 10; MOA-LX9N)...
Займер	18.07.2022, 01:03	Отклонено	Яндекс	NaN	NaN	NaN	NaN	Rezh	Десктоп	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.3...
Займер	22.07.2022, 01:01	Отклонено	Яндекс	NaN	NaN	NaN	NaN	Nachalovo	Мобильный телефон	Mozilla/5.0 (Linux; Android 10; JSN-L21) Apple...
Веббанкир	19.07.2022, 13:58	Одобрено	NaN	NaN	NaN	NaN	NaN	Yekaterinburg	Мобильный телефон	Mozilla/5.0 (iPhone; CPU iPhone OS 15_4_1 like...

Рис 1. : Пример данных из первого набора

Здесь каждая строка является транзакцией заёмщика в указанную в колонке “Оффер” организацию. Статус транзакции может быть одобренным либо отклонённым. Видно, что для разных транзакций каждая характеристика могла принимать разные значения. Некоторые характеристики содержали в себе большое количество пропусков, что указано на рис. 2.

Оффер	0.000000	Читают отзывы перед оформлением	99.722607	Версия браузера	27.323162
Время конверсии	0.000000	Город	0.000000	Модель устройства	11.003236
Статус	0.000000	Тип устройства	0.000000	Страна	0.000000
Поисковая система	40.221914	UserAgent	0.000000	Регион	0.000000
Плохая кредитная история	92.417938	ОС	0.000000	Тип соединения	0.000000
Возрастная группа	99.121590	Версия ОС	94.868239	Оператор	0.000000
Безработные	99.768840	Браузер	0.000000	IP	0.000000

Рис 2. : Процент пропусков в характеристиках

Мы считали, что характеристика с именем $feature$ является неинформативной, если условная вероятность одобрения транзакции

$$\mathbb{P}[\text{Статус} = \text{одобренно} \mid feature = value] \quad (1)$$

при любом значении $value$ этой характеристики оставалась примерно одинаковой. Чтобы найти эту вероятность, мы делили число тех одобренных транзакций, в которых $feature = value$, на число всех транзакций (считая и отклонённых), в которых $feature = value$. Такая оценка с увеличением числа транзакций будет сходиться к истинной вероятности (1).

Для примера на рис. 3 представлены условные вероятности одобрения займа для характеристики “Браузер”.

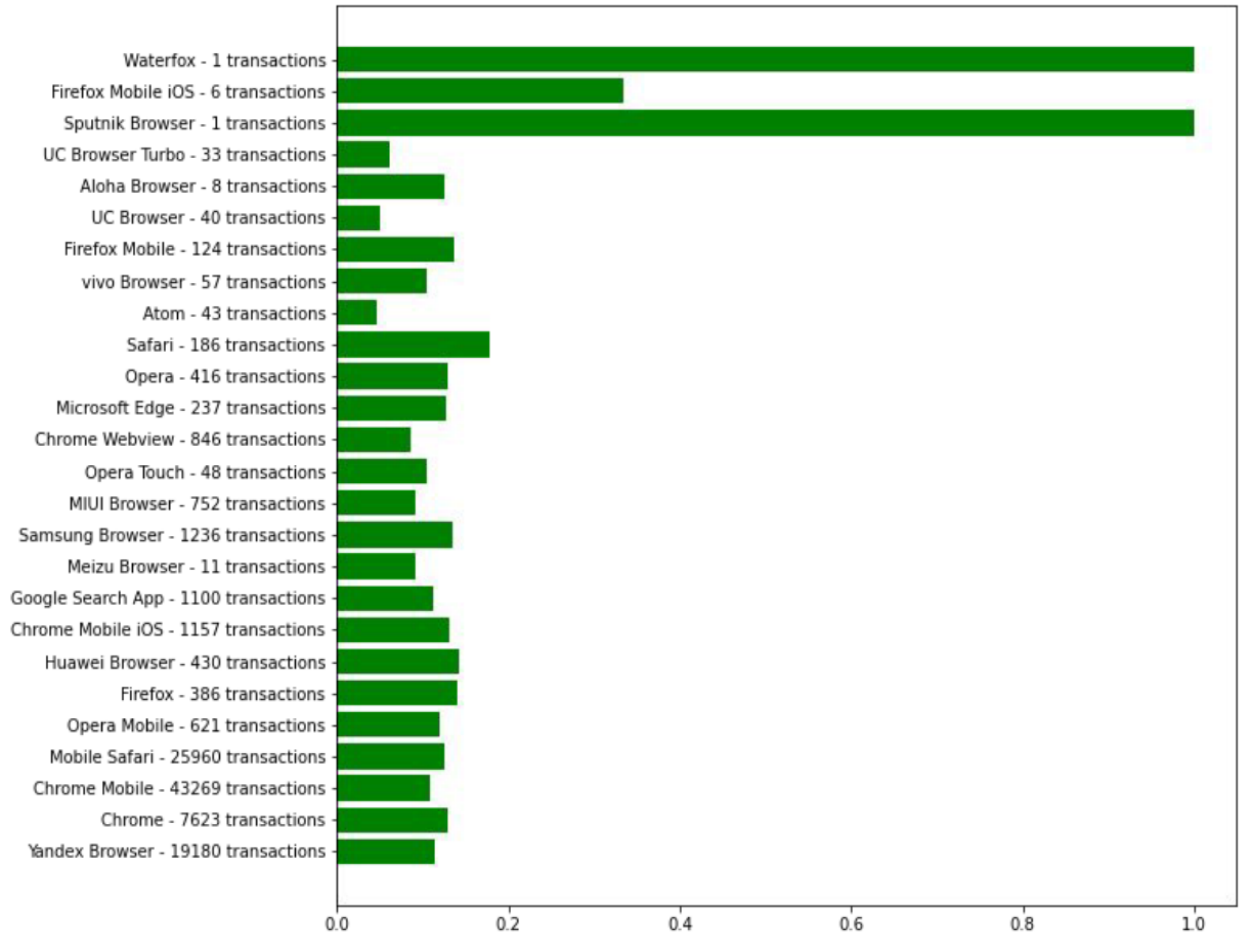

Рис 3. : Распределение условной вероятности в зависимости от характеристики “Браузер”

Из рис. 3 видно, что с увеличением числа транзакций вероятности удобрения займа для разных браузеров примерно одинаковы. Поэтому характеристика “Браузер” является неинформативной. По той же причине неинформативными характеристиками оказались: “Версия браузера”, “UserAgent”, “ОС”, “Версия ОС”, “Время конверсии”, “Читают отзывы перед оформлением”, “Город”, “Тип соединения”, “Оператор”. Их пришлось удалить. Кроме них мы удалили характеристики “Поисковая система”, “Плохая кредитная история”, “Возрастная группа”, “Безработные”, поскольку они содержали большое количество пропусков (рис. 2).

Информативными характеристиками оказались: “Оффер”, “Тип устройства”, “Модель устройства”, “Регион”. Мы оставили их. Для дальнейшей работы мы ввели оценки для характеристики “Оффер”, т.е. для организаций. Например, для организации “Займер” оценка имеет следующий вид:

$$\frac{\text{число одобренных транзакций в "Займер"}}{\text{число всех транзакций в "Займер"}} \cdot \ln(1 + \text{число всех транзакций в "Займер"}). \quad (2)$$

Первый множитель есть выборочная вероятность одобрения со стороны “Займер”, а второй - даёт вес этой вероятности. У нас было предположение, что чем выше эта оценка, тем выше должна быть вероятность одобрения. Потом мы смогли подтвердить это. Оценка (2) считалась для каждой организации.

4. Анализ второго набора данных.

Второй набор данных был гораздо меньше первого, поскольку содержал чуть более 2 тысяч транзакций. Но теперь каждая транзакция содержала в себе ещё и ответы на вопросы анкеты: сумма займа, срок займа, способ получения займа, доход заёмщика, наличие просрочек у заёмщика (рис. 4).

Оффер	Статус	Желаемая сумма	Желаемый срок	Способ получения	Ваш доход	Были ли просрочки	Тип устройства
Быстроденьги	lead	25000	15	Онлайн	10000	Нет	Мобильный телефон
Екапуста	lead	7000	6	Онлайн	30000	Нет	Мобильный телефон
ДжойМани	lead	18000	14	Онлайн	20000	Да	Мобильный телефон
Турбозайм	lead	13000	15	Онлайн	20000	Да	Мобильный телефон
Белка Кредит	lead	29000	8	Онлайн	15000	Да	Мобильный телефон
Екапуста	lead	12000	14	Онлайн	45000	Да	Мобильный телефон
Вива Деньги	lead	23000	15	Онлайн	25000	Да	Мобильный телефон

Рис 4. : Пример данных из второго набора

При работе с этими данными нами были предложена следующая характеристика

$$organizer\ estim = \frac{\text{Ваш доход}}{\text{Желаемый срок} \times \text{Желаемая сумма}} \times \text{Оценка организации.} \quad (3)$$

В формуле (3) мы использовали оценки, которые мы подсчитывали по формуле (2) для первого набора данных. При построении (3) мы заметили, что зачастую обращения, в которых большие желаемые срок и сумма, а также не высокий доход, имеют отклонённый статус. Верно было и обратное – когда доход, указанный в обращении, был высокий, а желаемые срок и сумма были невысокий, это обращение часто имело принятый статус. Поэтому мы предположили, что чем больше первый множитель в (3), тем больше будет вероятность одобрения. Второй множитель в (3) дополнительно повышал уверенность в одобрении займа. На рисунке 5 представлены распределения принятых и отклонённых транзакций в зависимости он характеристики (3).

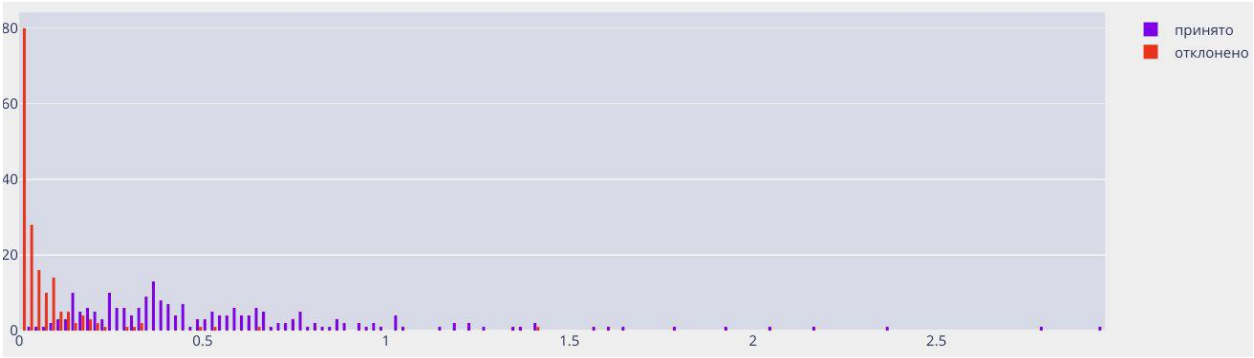

Рис. 5: Распределения транзакций в зависимости от характеристики (3)

Видно, что распределения отклонённых и принятых транзакций сильно различаются. Для отклонённых транзакций распределение близко к вырожденному около нуля, а для принятых – распределение преимущественно сосредоточено приблизительно правее 0.2. Причём если не использовать оценки (2) организаций в характеристике (3), то мы не получим такое явное разделение отклонённых и принятых транзакций.

5. Итоги работы.

В рамках проекта участники освоили основы работы с библиотеками Pandas, NumPy, Seaborn. Нашей команде удалось глубоко проанализировать полученные наборы данных, построить полезную характеристику, относительно которой распределения одобренных и отклонённых транзакций сильно отличаются. Поэтому по этой характеристике можно делать выводы о том, будет ли одобрен займ. Кроме того, мы смогли определить неинформативные характеристики, значения которых никак не влияют на одобрение займа.