

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК
ОСНОВНАЯ ОБРАЗОВАТЕЛЬНАЯ ПРОГРАММА
«ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА»

КУРСОВАЯ РАБОТА

Программный проект на тему:
«Кредитный скоринг. Сравнение линейных
моделей
с более сложными моделями машинного
обучения»

Выполнила студентка группы БПМИ194 3 курса,
Смирнова Анна Романовна

Руководитель КР:

Преподаватель Воробьева Мария Сергеевна

Оглавление

Аннотация	4
Abstract.....	5
Введение	6
Расшифровка признаков	8
Обработка данных	9
1. Предварительная обработка.....	9
2. Методы отбора переменных	10
2.1. Одномерные методы	10
2.2. Многомерные	13
3. Разведочный анализ данных	13
4. Information value.....	13
5. WOE	13
6. Information value.....	13
Метрики качества	13
1. F1	13
2. Precision	13
3. Recall	13
4. ROC AUC.....	13
5. Коэффициент GINI	13
6. Lift-Curve	13
Кодирование данных.....	13
1. LabelEncoder.....	13
2. OneHotEncodind	13
Древесные алгоритмы	14

1. Классическое дерево решений	14
2. Градиентный бустинг	14
3. RandomForest.....	14
4. XGBoost	14
5. DecisionTreeRegressor.....	14
6. CatBoost	14
7. LightGBM.....	14
8. Сравнительный анализ	14
Логистическая регрессия	14
1. На обычных переменных	14
2. На переменных с OneHotEncoding	14
3. На отмасштабированных переменных	14
4. На WOE переменных.....	14
5. RidgeClassifier	14
6. Сравнительный анализ	14
Заключение.....	15
Список литературы.....	16

Аннотация

Программный проект на тему: «Нейронные сети на табличных данных на примере задачи кредитного скоринга»

Выполнила: Смирнова А. Р.

Руководитель: Воробьева М. С.

В ходе данной курсовой работы был выполнен анализ качества существующих методов решения задачи кредитного скоринга. Главной целью являлось проанализировать и протестировать существующие методы решения данной задачи, после чего составить подробное описание каждого из методов и провести сравнительный анализ качества. Методы, которые были изучены: Ridge-регрессия, логистическая регрессия, метод опорных векторов, дерево решений, случайный лес.

Ключевые слова: кредитный скоринг, модель, метрика качества, модели машинного обучения, банковская сфера, кредит.

Abstract

Program course work: «Credit scoring. Linear models vs modern data science models»

Student: Smirnova A. R.

Teacher: Vorobeva M. S.

In the course work, the analysis of the quality of existing methods of solving the problem of credit scoring was carried out. The main objective was to analyze and test existing methods of solving this problem. The conclusion included a detailed description of each method and a comparative analysis of quality. Methods that have been studied: Ridge-regression, logistic regression, support vector machine, decision tree, random forest.

Keywords: credit scoring, neural network, machine learning models, quality metric, banking industry, loan.

Введение

Задача кредитного скоринга возникает в ситуации, когда банку необходимо принять решение о выдаче или отказе по кредиту, при этом принимая во внимание множество несвязанных факторов. Одним из решений данной ситуации является субъективное заключение кредитного эксперта, однако человеческий фактор не всегда позволяет учесть все входные данные, потому что с течением времени количество факторов, влияющих на принятие решение о выдаче кредита, растет.

Банковская сфера активно расширяется, и каждой компании необходимо внедрять качественные автоматизированные системы, которые помогут эффективно обеспечивать контроль работы бизнес-процессов. Решить кредитный вопрос активно помогают системы кредитного скоринга, такие как, например, нейронные сети на табличных данных, или другие модели классификации, с помощью которых можно оценить большой набор признаков и учесть предыдущий опыт работы с кредиторами.

Главной **целью** этой работы является самостоятельно применить на табличных данных существующие методы решения задачи кредитного скоринга, после чего составить подробное описание каждого из методов и провести сравнительный анализ качества. **Актуальность** работы подтверждается постоянным совершенствованием банковской системы, и растущей необходимостью работать с большими объемами данных. Численность населения непременно растет, в связи с чем увеличивается и количество предлагаемых банками услуг, в том числе различного рода кредитов, отличных от привычного понимания. Например, появлением

лизинга¹ или популярности микрозаймов², которые так же нуждаются в анализе платежеспособности своих клиентов.

Основные **задачи**, которые предстоит выполнить в ходе работы:

1. Провести предварительную обработку имеющихся данных.
2. Реализовать различные методы машинного обучения: в контексте моей работы это различные архитектуры нейронных сетей и древесные алгоритмы, которые было решено использовать для сравнения.
3. Провести сравнительный анализ полученных результатов через метрики качества: F1, ROC AUC, Precision, GINI, Ассигасу и другие.
4. Формулировать предложения о том, как можно внедрить данные системы для эффективной работы в банковской сфере.

¹ Лизинг — это долгосрочная аренда определенных объектов собственности (оборудование, машины, сооружения) с погашением задолженности в течение нескольких лет.

² Микрозаймы — это кредиты на небольшие денежные суммы, которые обычно даются на короткий период.

Расшифровка признаков

пока просто скопировала из ноутбука

loan_amnt — запрашиваемая у банка сумма для кредита

funded_amnt — накопительная сумма

funded_amnt_inv — накопительная сумма для инвестора

term — на какой промежуток выдан кредит

int_rate — процентная ставка

installment — первоначальный взнос

grade — оценка кредитного риска

emp_title — профессия

emp_length — опыт работы

home_ownership — тип жилья

annual_inc — годовые доходы

verification_status — статус верификации

issue_d — месяц, в который получено финансирование

purpose — цель

addr_state — кто выдает кредит (гос. сторона)

dti — это процент ежемесячного валового дохода потребителя, который идет на выплату долгов

delinq_2yrs — просрочка платежей более чем на 30 дней за последние два года

earliest_cr_line — месяц открытия самой ранней кредитной истории

inq_last_6mths — количество обращений кредитора (то есть Банка) в бюро кредитных историй, если клиент ходит по разным банкам и просит кредиты, то это говорит, что этому клиенту либо отказывают, либо он решил набрать везде кредитов. Как правило это негативно влияет

open_acc — это количество открытых кредитов

revol_bal — общий оборотный остаток

revol_util — доля утилизации. Клиенту выдали кредитную карту с лимитом 100 рублей, он уже потратил 40 рублей, **revol_util** будет равен $40/100 = 0.4$

total_acc — видимо все кредиты и активные, и закрытые
out_prncp — оставшаяся непогашенная сумма
total_pymnt — полученные на сегодня платежи
loan_status — текущий статус кредита (закрыт не закрыт)
risk — выдаем не выдаем

Обработка данных

Предварительная обработка данных необходима перед решением любой задачи машинного обучения. При сборе данных для обучения модели могут быть использованы различные источники, что может послужить наличию выбросов, пропущенных данных, расхождения формата и настоящего значения признака и другого.

Зачастую предобработка данных занимает больше всего времени, однако тщательное и подробное изучение имеющегося набора признаков — залог успешного решения задачи, поскольку более качественные данные помогают получить более совершенную модель. Для того чтобы принять решение, как поступать с той или иной проблемой, существует множество методов. Мы посмотрим на подробное описание каждой из них, и постепенно применим наиболее подходящие к нашему набору данных.

1. Предварительная обработка

Перед тем как переходить к точечным методам, стоит провести общую обработку данных. Она будет включать в себя следующие шаги:

Поскольку в моей задаче достаточное количество таргета обоих значений, я буду удалять любые признаки с более чем 85% пропущенных значений, не боясь потерять ценные данные.

Удалению также подлежат признаки, у которых только одно уникальное значение в совокупности с пропущенными данными. Таких колонок в моем дата сете обнаружено не было.

Удалить признаки, значения которых повторяют друг друга, чтобы избежать мультиколлинеарности³. В нашем наборе признаков, точно повторяющих другие нет, но есть те, которые включают в себя значения другого признака. Например: *funded_amnt* и *funded_amnt_inv*, распределение которых достаточно близки. В такой ситуации имеет смысл оставить только один признак, для нас это *funded_amnt_inv*, так как он имеет бóльший объем данных. Можно это увидеть так же на рисунке 1.1, где коэффициент Пирсона между этими признаками равен единице.

После удаления основных пропусков и повторений, стоит рассмотреть каждый признак в отдельности. Данный подход не всегда является оптимальным, поскольку признаков может быть слишком много, но в нашем случае это поможет построить более качественную модель, так как количество колонок не велико. При рассмотрении каждого признака можно обратиться к нескольким методам.

2. Методы отбора переменных

2.1. Одномерные методы

Одномерными называются методы, отражающие зависимость между конкретным признаком и целевой переменной.

Самый популярный пример данного типа – корреляция. Корреляция – это показатель степени зависимости между двумя переменными. Интервал значения корреляции может быть от -1 до 1 . Положительная корреляция говорит о том, что при увеличении значения одной переменной будет увеличиваться и другая, и наоборот. При работе с данными чаще всего используют корреляции *Спирмена*, *Кендалла* и *Пирсона*.

³ Мультиколлинеарность — случай, при котором наблюдается высокая корреляция между признаками. Для проверки на мультиколлинеарность можно посмотреть на коэффициент Пирсона между ними.

При выборе коэффициента стоит учитывать, к какому типу принадлежит исследуемая переменная. Например, при нормальном распределении переменной стоит смотреть на значение коэффициента Пирсона, поскольку она подходит для непрерывных метрических переменных, а в случае ненормального распределения – Спирмена, как мере, которая работает с признаками, измеренными в ранговой шкале. Кендалла же необходимо использовать для рассмотрения связи двух переменных, представленных двумя порядковыми шкалами.

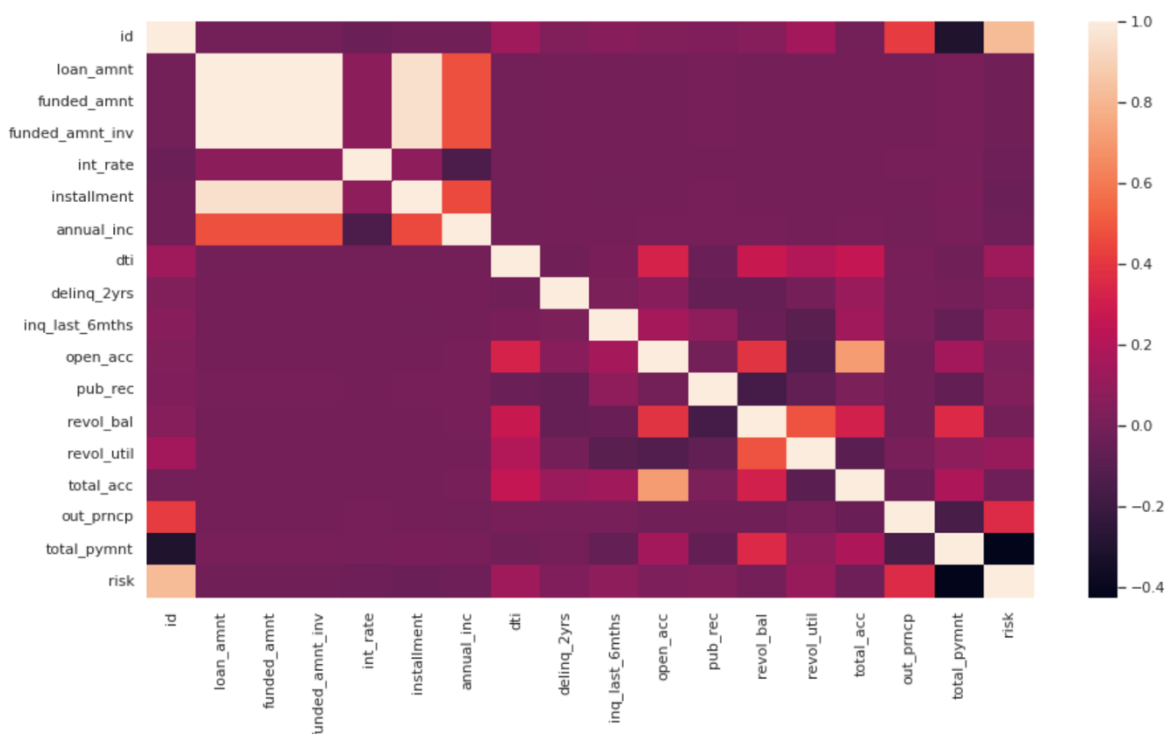


Рис 1.1 Корреляция Спирмена с целевой переменной перед обработкой данных

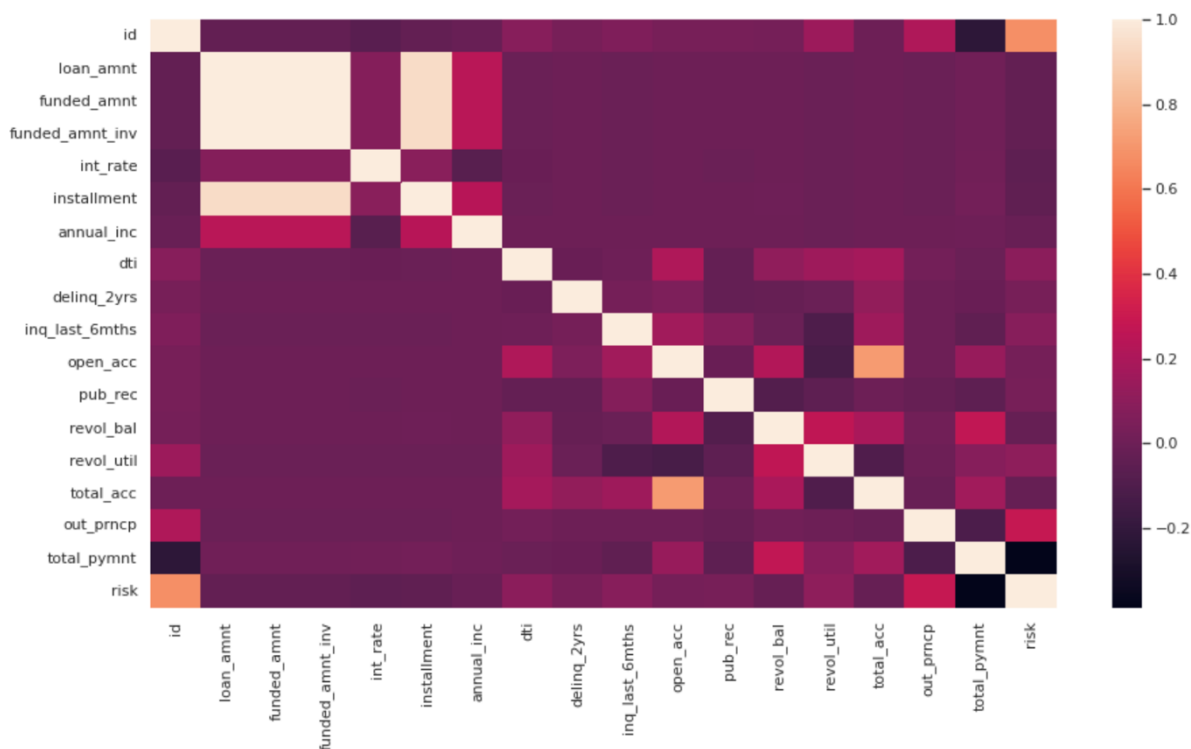


Рис 1.1 Корреляция Пирсона с целевой переменной перед обработкой данных

Другой известной одномерной характеристикой является *Хи-квадрат* – статистический критерий, который используется для анализа связи между двумя категориальными переменными. Данный метод используется с самого зарождения задачи кредитного скоринга. В 1941 г. в «National Bureau of Economic Research»⁴ Дэвид Дюран опубликовал свое исследование, в котором использовал более семи тысяч «хороших» и «плохих» кредитных историй. Тогда хи-квадрат был использован для поиска отличительных черт, которые наиболее остро делили два типа кредитных историй, после чего Дэвидом был разработан индекс эффективности, предназначенный для демонстрации того, насколько эффективна данная характеристика для дифференциации степени риска среди заявителей на кредиты.

⁴ Исследование «Risk Elements in Consumer Instalment Financing»

F-меры

2.2 Многомерные (совокупность переменных)

3. Разведочный анализ данных

4. Information value

5. WOE

6. Information value

6.1 Boruta

Метрики качества

Какой смысл / какие подходят для нашей задачи

1. F1

2. Precision

3. Recall

4. ROC AUC

5. Коэффициент GINI

6. Lift-Curve

Кодирование данных

1. LabelEncoder

2. OneHotEncodind

Древесные алгоритмы

Чем они отличаются и какие предоставляют улучшения

1. Классическое дерево решений
2. Градиентный бустинг
3. RandomForest
4. XGBoost
5. DecisionTreeRegressor
6. CatBoost
7. LightGBM
8. Сравнительный анализ

Логистическая регрессия

1. На обычных переменных
2. На переменных с OneHotEncoding
3. На отмасштабированных переменных
(минус среднее / среднее квадратическое, например)
4. На WOE переменных
5. RidgeClassifier
6. Сравнительный анализ

Заключение

Список литературы

- Ahamed, S. (б.д.). *Machine learning algorithms can help us to estimate the risk of a financial decision*. Получено из Towards Data Science: <https://towardsdatascience.com/financial-data-analysis-80ba39149126>
- Complex System Modelling in Engineering Under Industry*. (22 10 2021 г.). Получено из Hindawi: <https://www.hindawi.com/journals/complexity/2021/9222617/>
- Essche, P. O. (2018). *Solvay Brussels School*. Получено из Big Data for Credit Scoring: towards the End of Discrimination on the Credit Market? Evidence from Lending Club.
- Python_Credit_Scoring (ML)_Elastic net regression*. (б.д.). Получено из kaggle: <https://www.kaggle.com/code/aashofteh/python-credit-scoring-ml-elastic-net-regression/data>
- Risk Elements in Consumer Instalment Financing*. (б.д.). Получено из National Bureau of Economic Research: <https://www.nber.org/books-and-chapters/risk-elements-consumer-instalment-financing-technical-edition>
- Отбор признаков в задачах машинного обучения*. (б.д.). Получено из Хабр: <https://habr.com/ru/post/550978/>