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

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, and 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 больший 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

лизинга[1] или популярности микрозаймов[2], 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, Accuracy 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

loan_amnt – Запрашиваемая у банка сумма для кредита.

funded_amnt — Общая сумма обязательств по кредиту на данный момент.

funded_amnt_inv — Общая сумма обязательств инвесторов.

term — Количество платежей по запрошенной сумме.

int_rate — Процентная ставка кредита.

installment — Размер первоначального взноса.

grade — Оценка кредитного риска.

emp_title — Должность, представленная заемщиком при получении кредита.

emp_length — Трудовой опыт в годах.

home_ownership — Статус собственности жилья на момент открытия кредита.

annual_inc — Годовой доход.

verification_status — Статус верификации.

issue_d — Месяц, в который получено финансирование.

purpose — Цель взятия кредита.

addr_state — Государство, указанное в заявке на получение кредита.

dti — Процент ежемесячного валового дохода потребителя, который идет на выплату долгов.

delinq_2yrs — Количество просроченных платежей более чем на 30 дней.

earliest_cr_line — Месяц открытия самой ранней кредитной истории.

inq_last_6mths — Количество обращений кредитора в бюро кредитных историй.

open_acc — Это количество открытых кредитов в данный момент.

revol_bal — Общий кредитный оборотный остаток.

revol_util — Доля утилизации кредита

total_acc — Общее количество активных и закрытых кредитов.

out_prncp — Оставшаяся непогашенная сумма.

total_pymnt — Выплаты, полученные на сегодняшний день.

loan_status — Текущий статус кредита.

risk — Дефолт или не дефолт.

1. 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. Разведочный анализ данных

Exploratory Data Analysis (EDA) – это важнейшая ступень подготовки к построению модели. Этот этап обычно занимает больше всего времени, потому что мы вручную исследуем каждый признак на предмет значимости для решения нашей задачи.

Для рассмотрения категориальных признаков основной задачей является посмотреть на уникальные значения и их интерпретируемость. Например, в изначальном наборе данных у нас есть переменная *loan_status,* которая принимает такие значения: Fully Paid, Charged Off, Default, In Grace Period, Late (16–30 days), Late (31–120 days). Если мы знаем каждое из значений, то можем более точно оценить, какие категориальные переменные подойдут нам лучше. Например, здесь нам важнее оставить только два из них – Fully Paid, Charged Off, поскольку нас интересует только два сценария.

Для числовых признаков ситуация обстоит иначе: здесь нам сложно интерпретировать значения, поэтому важнее будет посмотреть на их математические характеристики. Например, на график распределения, который поможет применить правильные модели отбора переменных в

дальнейшем, на «ящик с усами» для проверки наличия выбросов, наличие отрицательных значений.

2. 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 мультиколлинеарности[3]. 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.

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

3. Методы отбора переменных

3.1. 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*.

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.

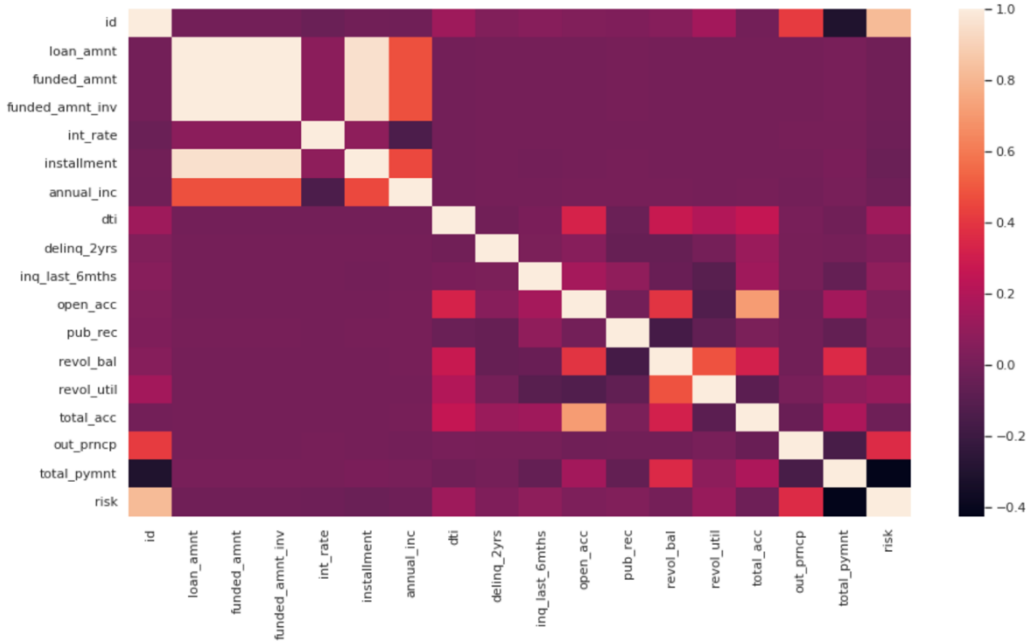

Рис 1.1 Корреляция Спирмена с целевой переменной перед обработкой данных

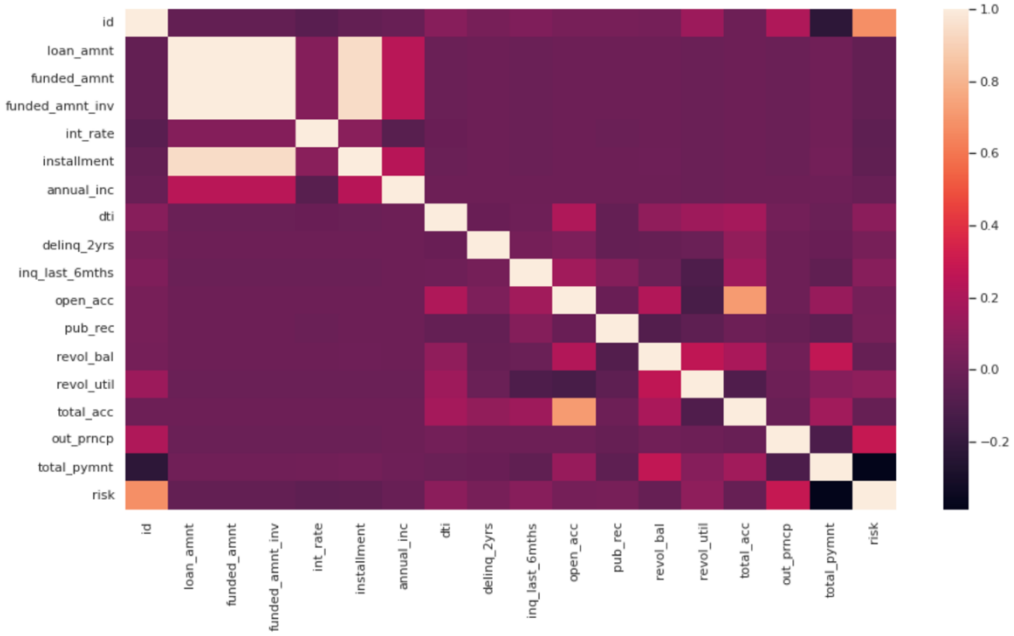

Рис 1.2 Корреляция Пирсона с целевой переменной перед обработкой данных.

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.

Для примера можем принять за нулевую гипотезу, что переменные «Оценка кредитного риска» и «дефолт» не зависят друг от друга. Для этого реализовываем тест Chi-Square и получаем P-значение равным 0.03267 < 0.05,

[4] Исследование «Risk Elements in Consumer Instalment Financing»

что говорит о необходимости отвергнуть нулевую гипотезу и принять, что эти признаки зависят друг от друга.

2.2 Многомерные методы

Если с одномерными переменными мы имеем четкий и структурированный набор методов, то в случае оценки эффективности нескольких переменных мы чаще пользуемся метриками качества, результаты которых сравниваем при разных наборах переменных.

Создание новых признаков может значительно улучшить качество работы модели, ровно как и наоборот. Ситуаций, в которых мы считаем полученную комбинацию признаков или новый созданный признак удачным, может быть несколько, как например:

- Повышение точности предсказаний. Измеряется с помощью различных метрик, о которых подробнее поговорим в следующей главе.
- Более легкая интерпретируемость результатов.

Стоит так же уточнить, что в некоторых ситуациях признаки, не содержащие значительной информации, могут быть полезны для создания новых, более качественных признаков. Это особенно стоит брать во внимание, когда общее количество переменных небольшое и «выбрасывать» из рассмотрения целый столбец может быть неэффективно.

Посмотрим на несколько примеров многомерных методов. Один из самых наглядных – построить *диаграмму рассеяния*. С ее помощью можно наглядно отследить, наблюдается ли корреляция между двумя независимыми признаками, если мы заранее можем предположить сильную связь между ними. A – наивысший уровень кредитоспособности, G – наименьший.

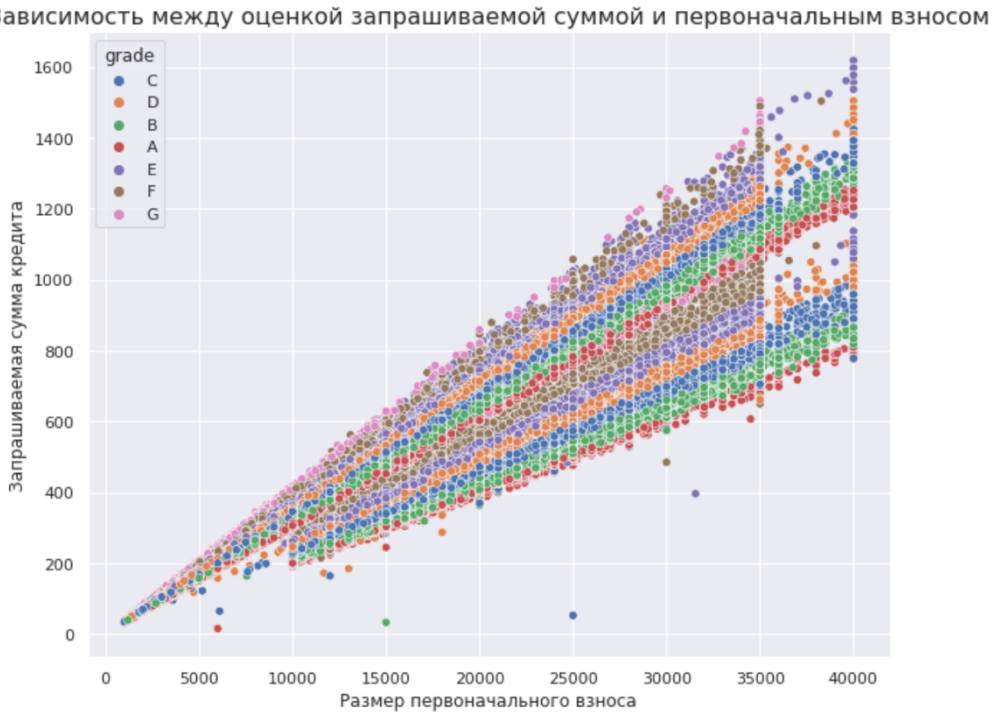

Рис 1.3 Диаграмма рассеяния первоначального взноса и запрашиваемой суммы кредита для разных оценок кредитного риска.

Здесь не хватает практической части, надо попридумывать новых признаков, чтобы посмотреть на всякие зависимости в главе с метриками

4. WOE и IV

При решении задачи бинарной классификации мы часто имеем дело со множеством категориальных признаков, оценить вклад которых бывает сложно. Для решения этой проблемы разберем две концепции, которые широко применимы при решении задачи кредитного скоринга: «Weight of evidence» и «Information value».

$$WOE = \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

Рисунок 1.4 Вычисление WOE и IV, где Even – дефолт, Non Event – его отсутствие

Вычисление полезности какого-либо признака через WOE преобразование очень близко к вычислению обычной байесовской вероятности, мы смотрим логарифм от соотношения процента «хороших клиентов» к рассматриваемой характеристикой к количеству «плохих». После вычисления значения *weight of evidence* у каждого признака, можно попробовать объединить столбцы с близким значением WOE. Это связано с тем, что эти признаки могут попросту одинаково влиять на целевую переменную, а значит нет необходимости оставлять их обе сепарировано, чтобы не вызвать переобучения.

Некоторые плюсы использования WOE:

- Самостоятельно обрабатывает пропущенные значения
- Работает с выбросами
- Итоговое значение это логарифм распределения, что согласуется со значением логистической регрессии

Information value – это один из наиболее простых способов оценки значимости признака. После выполнения WOE преобразования можно вычислить «коэффициент полезности», после которого принять решение полезен ли признак в нашей ситуации.

Information Value	Variable Predictiveness
Less than 0.02	Not useful for prediction
0.02 to 0.1	Weak predictive Power
0.1 to 0.3	Medium predictive Power
0.3 to 0.5	Strong predictive Power
>0.5	Suspicious Predictive Power

Рисунок 1.4 Согласно «Credit Risk Scorecards»[5], оценка IV в задаче кредитного скоринга может трактоваться следующими значениями.

У WOE и IV пока нет практической части

[5] Книга «Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring», автор: Naeem Siddiqi

2. Метрики качества

После обработки всех данных и построения модели, важно оценить ее эффективность. Посмотрим на распространенные подходы к измерению качества моделей, и определим, какие из них подойдут к нашей задаче.

Глобально метрики качества бинарных классификаторов делятся на «пороговые» (single-threshold) и «не зависящие от порога» (threshold-free). Чтобы разделить значения на две категории, необходимо найти «порог», по которому это разделение будет проведено – именно это отличает два этих типа друг от друга. Не зависящие от порога метрики проводят разделение разными способами, которые мы рассмотрим ниже. Считается, что в задаче кредитного скоринга исторически наиболее часто используются именно threshold-free метрики.

1. Accuracy

Наиболее очевидной метрикой при обучении в задаче классификации является доля правильных ответов. Она является пороговой, и сложно интерпретирует полезность полученной модели (особенно на несбалансированных наборах данных), потому что процент положительных ответов не всегда содержит в себе какую-то полезную информацию. Однако, точность является частью более сложных метрик, и несмотря на то, что ее я использовать в своих решениях не буду, важно было вспомнить ее значение.

2. Precision и Recall

Более информативными являются две следующих метрики: *точность* (precision) и *полнота* (recall). Для того чтобы их рассчитать, нам необходимо обратиться к матрице ошибок. Матрица ошибок – это способ разделения всех результатов на четыре категории, в зависимости от предсказания модели и реального ответа: True Positive, False Positive, True Negative, False Negative. Из этих данных точность и полнота высчитываются следующим образом:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Полнота отражает долю верно выделенных классификатором положительных объектов. Точность отражает долю положительно выделенных объектов, которые действительно являются положительными. Однако, обе метрики стоит особенно осторожно применять на несбалансированных данных. В задаче кредитного скоринга чаще используется комбинация этих метрик, которая создает новый способ оценки ошибки: F меру.

3. F-мера

F-мера (или F1) является средним гармоническим точности и полноты. Эта метрика идеально подходит, если мы хотим сравнить точность предсказания используя результаты работы разных моделей. Что отличает эту метрику от трех предыдущих: она полезна даже на несбалансированных наборах данных. Однако, ее сложнее интерпретировать, чем, например, accuracy.

4. ROC-AUC

Чтобы перейти к описанию данной метрики, нужно ввести еще два понятия: *False Positive Rate* и *True Positive Rate*. Первая отражает долю неверно определенных дефолтных заемщиков, вторая – долю верно определенных. Обе метрики формируются из обозначенной ранее матрицы ошибок:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}; \qquad \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Рассмотрим координатную плоскость, где двумя координатами будут FPR и TPR. Значения координат будут соответстенно от 0 до 1. Площадь под кривой, полученной из всех комбинаций данных значений, и будет являться данной интегральной метрикой. Таким образом, через количественное значение этой метрики, можно оценить качество классификатора, зная, например что ROC-AUC случайной модели равно 1/2.

5. 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

   Тут я уже поделала практическую часть, но не написала в отчет

2. На переменных с OneHotEncoding

   Тут я уже чуть-чуть поделала практическую часть, но не написала в отчет

			Metrics	Parameters		
	↓ Start Time	Duration	F1	Average of F1	Encoding type	Scaling type
	11 seconds ago	8.7s	0.869	weighted	OneHotEncoder	StandardScaler
	36 seconds ago	10.7s	0.589	weighted	OneHotEncoder	MinMaxScaler
	1 minute ago	8.9s	0.596	weighted	OneHotEncoder	StandardScaler
	4 minutes ago	12.0s	0.598	weighted	OneHotEncoder	StandardScaler
	6 minutes ago	10.8s	0.041	binary	OneHotEncoder	StandardScaler
	6 minutes ago	11.8s	0.434	macro	OneHotEncoder	StandardScaler
	7 minutes ago	9.5s	0.434	-	OneHotEncoder	StandardScaler

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/