Контрольное задание

Сусина Алиса, Барышев Григорий

November 2023

Содержание

1 Вступление **2**

2 Разведочный анализ данных (EDA) **2**

3 Правка данных **6**

4 Линейная регрессия **7**

4.1 Теория 7

4.2 Простейшая модель линейной регресии 8

4.3 Усложненная модель линейной регресии 9

5 Модель классификации **9**

6 Заключение **10**

1 Вступление

В данном задании нам было необходимо выбрать датасет, провести предварительный анализ данных, построить несколько моделей и описать соответсвующие результаты. Наш выбор остановился на датасете показывающем кредитные заявки и решение банка об одобрение или неодобрении заявки.

2 Разведочный анализ данных (EDA)

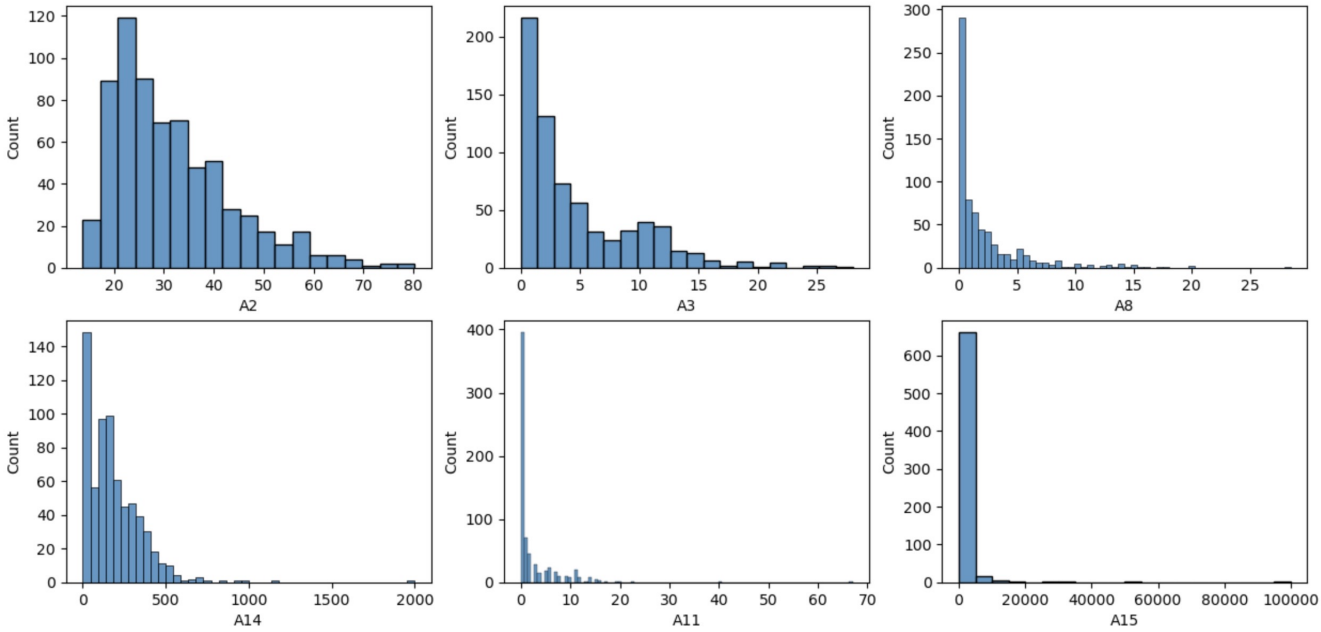
Для начала было необходимо получить общую информацию о данных: наличение пропущенных значений, тип переменных, название столбцов.

	name	role	type	demographic	description	units	missing_values
0	A16	Target	Categorical	None	None	None	no
1	A15	Feature	Continuous	None	None	None	no
2	A14	Feature	Continuous	None	None	None	yes
3	A13	Feature	Categorical	None	None	None	no
4	A12	Feature	Categorical	None	None	None	no
5	A11	Feature	Continuous	None	None	None	no
6	A10	Feature	Categorical	None	None	None	no
7	A9	Feature	Categorical	None	None	None	no
8	A8	Feature	Continuous	None	None	None	no
9	A7	Feature	Categorical	None	None	None	yes
10	A6	Feature	Categorical	None	None	None	yes
11	A5	Feature	Categorical	None	None	None	yes
12	A4	Feature	Categorical	None	None	None	yes
13	A3	Feature	Continuous	None	None	None	no
14	A2	Feature	Continuous	None	None	None	yes
15	A1	Feature	Categorical	None	None	None	yes

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
0	b	30.83	0.000	u	g	w	v	1.25	t	t	1	f	g	202.0	0	+
1	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	43.0	560	+
2	a	24.50	0.500	u	g	q	h	1.50	t	f	0	f	g	280.0	824	+
3	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	100.0	3	+
4	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	120.0	0	+
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
685	b	21.08	10.085	y	p	e	h	1.25	f	f	0	f	g	260.0	0	-
686	a	22.67	0.750	u	g	c	v	2.00	f	t	2	t	g	200.0	394	-
687	a	25.25	13.500	y	p	ff	ff	2.00	f	t	1	t	g	200.0	1	-
688	b	17.92	0.205	u	g	aa	v	0.04	f	f	0	f	g	280.0	750	-
689	b	35.00	3.375	u	g	c	h	8.29	f	f	0	t	g	0.0	0	-

На этом этапе мы столкнулись с проблемой, информация о заявках была зашифрованна в целях конфиденциальности, поэтому было необходимо получить более детальную информацию о данных, чтобы сформировать полную картину того с чем мы имеет дело. На этом этапе нами было постренно несколько различных графиков.

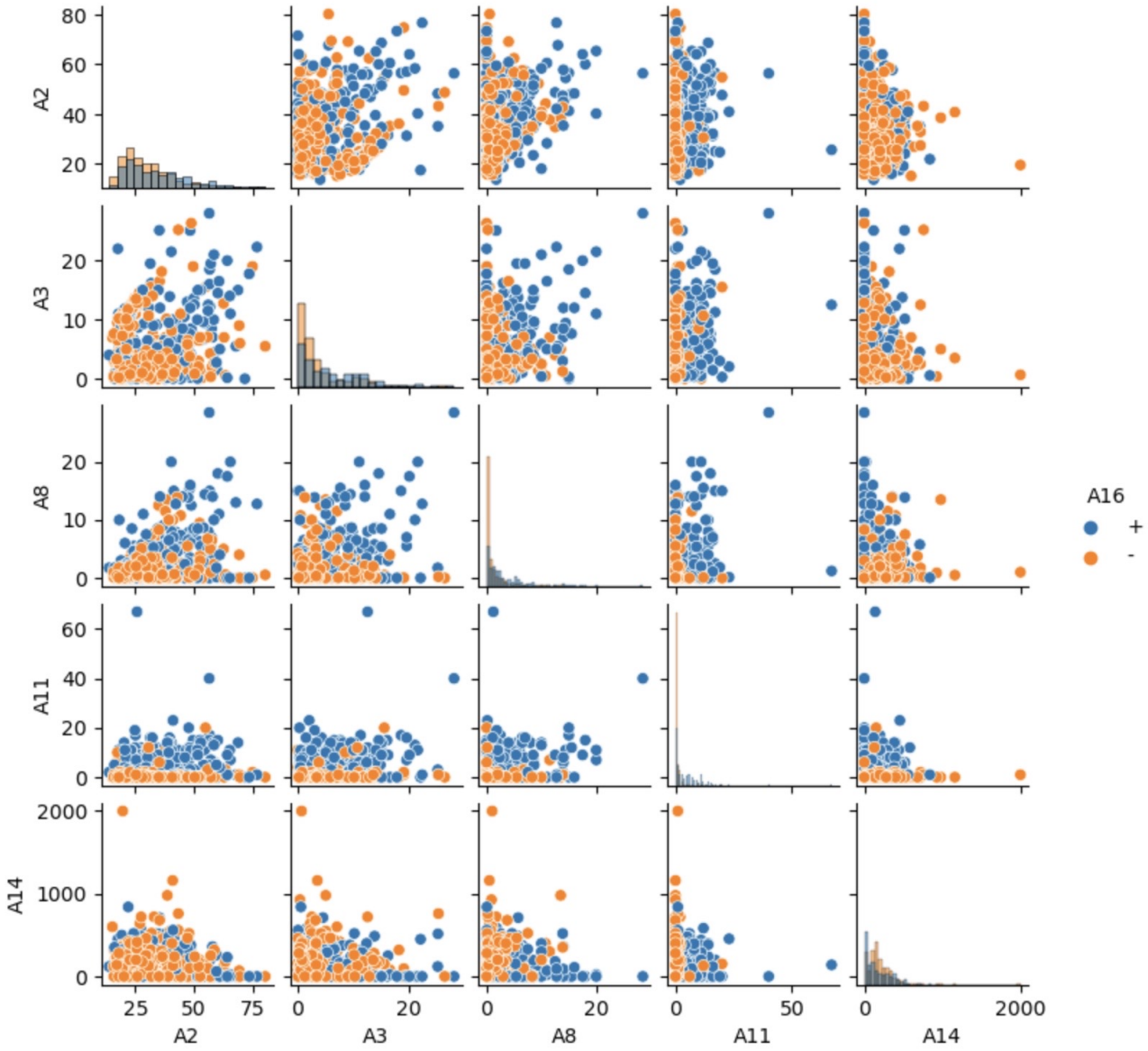
1. Гистограмма диапазонов числовых данных



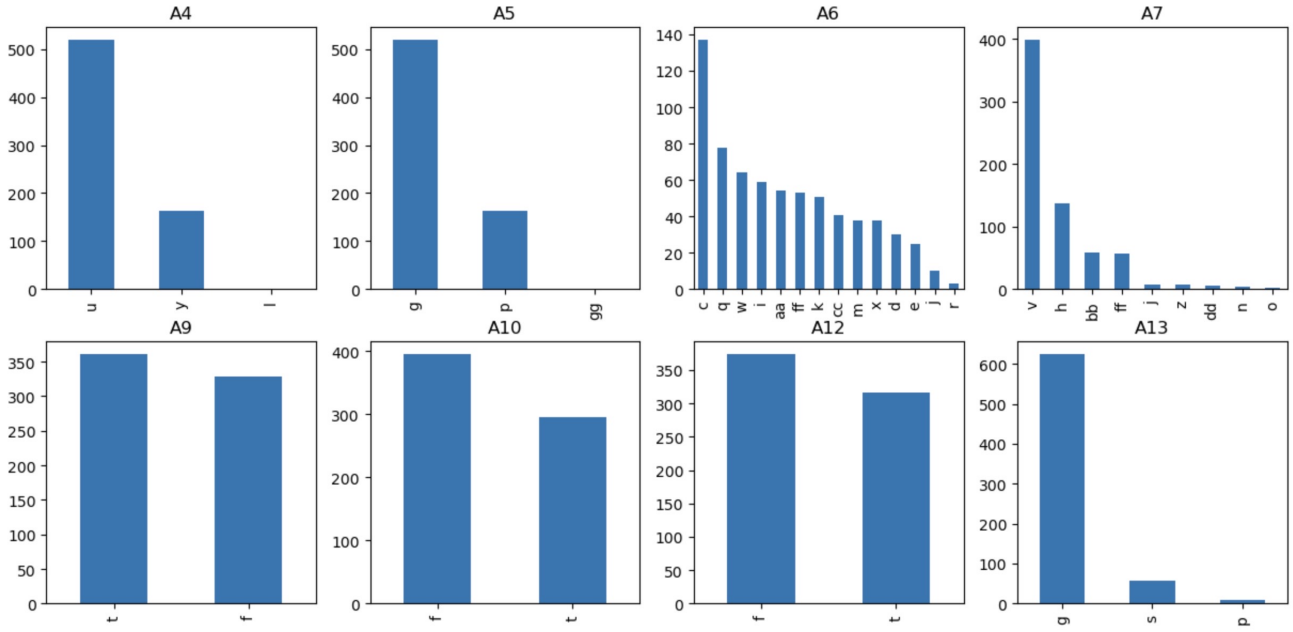
2. Информация о числовых значениях по колонкам

	count	mean	std	min	25%	50%	75%	max
A2	678.0	31.568171	11.957862	13.75	22.6025	28.46	38.2300	80.25
A3	690.0	4.758725	4.978163	0.00	1.0000	2.75	7.2075	28.00
A8	690.0	2.223406	3.346513	0.00	0.1650	1.00	2.6250	28.50
A11	690.0	2.400000	4.862940	0.00	0.0000	0.00	3.0000	67.00
A14	677.0	184.014771	173.806768	0.00	75.0000	160.00	276.0000	2000.00
A15	690.0	1017.385507	5210.102598	0.00	0.0000	5.00	395.5000	100000.00

3. Связь между собой различных признаков, с распределением по одобрению



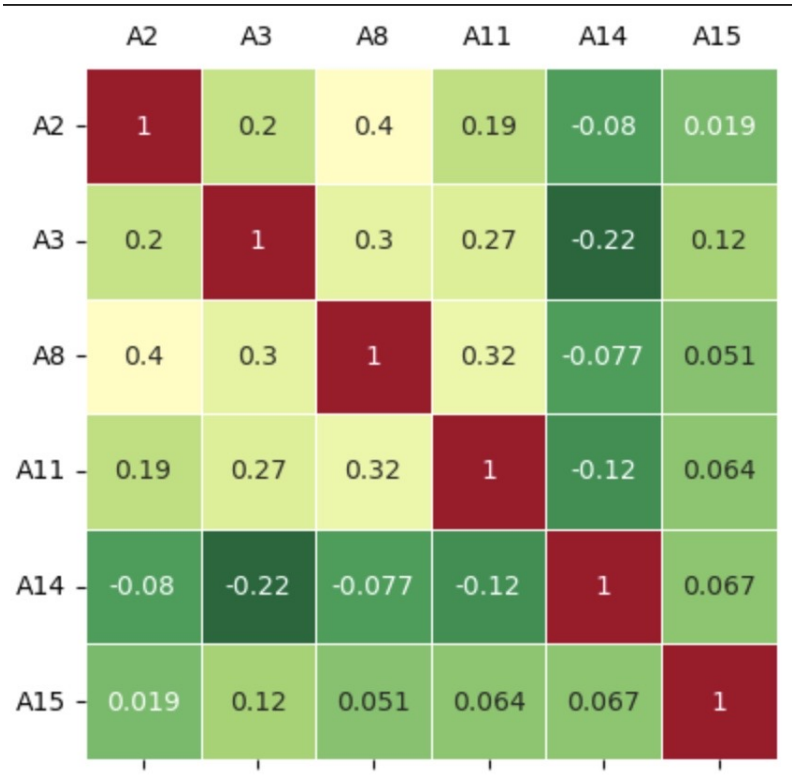
Распределение значений столбцов с строками(буквами)



Тепловая карта чтобы показать корреляцию между данными

Корреляция — математический показатель, по которому можно судить, есть ли статистическая взаимосвязь между случайными величинами. Если такая связь существует, изменения значений одной величины влияют на другую.

Коэффициент корреляции показывает, насколько велика взаимосвязь. Обозначается символами R или r и может принимать значения от -1 до 1 включительно.



3 Правка данных

В ходе разведочного анализа было замечено, что исходные данные неполные и находятся в формате неблагоприятном для проведения анализа. Поэтому перед построением моделей необходимо было обработать данные и привести их к удобному для работы виду.

Во-первых, данные были заполнены не полностью, имели столбцы с пропущенными значениями. Существует много решений этой проблемы таких как:

- Удаление строк или столбцов
- Заполнение средними значениями или медианой
- Замену пропущенных значений нулями и добавление специального атрибута-индикатора
- Метод заполнения вперед (forward fill) или назад (backward fill),
- Восстановление пропусков на основе регрессионных моделей

Чтобы оценить серьезность проблемы и найти оптимальное решение был проведен анализ данных, который показал что почти половина столбцов имеют пропуски, при этом процент пропусков в столбцах относительно невысок

```
Процент пропущенных данных для столбца A1 : 0.018
Процент пропущенных данных для столбца A2 : 0.018
Процент пропущенных данных для столбца A4 : 0.009
Процент пропущенных данных для столбца A5 : 0.009
Процент пропущенных данных для столбца A6 : 0.013
Процент пропущенных данных для столбца A7 : 0.013
Процент пропущенных данных для столбца A14 : 0.019
```

С учетом этого, были сделаны выводы, что из за большого колличество столбцов с пропусками, удаление столбцов или замена пропущенных значений нулями слишком сильно бы попортили дату, а такие методы как умное заполнение просто излишни. В то же время методы как forward fill и backward fill относительно просты и хорошо сохраняют структуру исходных данных, поэтому было решено имплементировать именно их, а более конкретно метод data.ffill(), который заполняют все пропущенные в датасете данные на основе предыдущих.

Вторая проблема данных состояла в их неудобном для анализа виде. Многие столбцы имели тип object, с которым крайне неудобно работать, так как большинство анализа строится на численных значениях. Было проведен перевод данных их типа object в численный тип с помощью метода fit_transform из библиотеки sklearn.preprocessing

В итоге был получен датафрейм без пропущенных значений и состоящий полностью из числовых значений на котором далее были построены модели.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16
0	1	30.83	0.000	1	0	12	7	1.25	1	1	1	0	0	202.0	0	0
1	0	58.67	4.460	1	0	10	3	3.04	1	1	6	0	0	43.0	560	0
2	0	24.50	0.500	1	0	10	3	1.50	1	0	0	0	0	280.0	824	0
3	1	27.83	1.540	1	0	12	7	3.75	1	1	5	1	0	100.0	3	0
4	1	20.17	5.625	1	0	12	7	1.71	1	0	0	0	2	120.0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
685	1	21.08	10.085	2	2	4	3	1.25	0	0	0	0	0	260.0	0	1
686	0	22.67	0.750	1	0	1	7	2.00	0	1	2	1	0	200.0	394	1
687	0	25.25	13.500	2	2	5	2	2.00	0	1	1	1	0	200.0	1	1
688	1	17.92	0.205	1	0	0	7	0.04	0	0	0	0	0	280.0	750	1
689	1	35.00	3.375	1	0	1	3	8.29	0	0	0	1	0	0.0	0	1

4 Линейная регрессия

4.1 Теория

Общая постановка задачи линейной регрессии. Есть k переменных $x_1, \ldots, x_k$ ("регрессоров"), через которые мы хотим выразить "объясняемую переменную"y:

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_k x_k.$$

Значения всех переменных мы измерены n раз. Подставим эти данные в предыдущее равенство:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = a_1 \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + a_2 \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \ldots + a_k \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{pmatrix}$$

(здесь x_{ij} — это значение j-го признака на i-м измерении). Это удобно переписать в матричном виде:

$$\begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

или коротко $Xa = y$. Поскольку на практике эта система уравнений зачастую не имеет решения (ибо зависимости в жизни редко бывают действительно линейными), методом наименьших квадратов ищется псевдорешение.

Самый популярный метод оценки качества модели - с помощью **среднеквадратичной ошибки** (mean square error). Она равна

$$\frac{1}{n}|X\hat{a} - y|^2 = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{a}_1 x_{i1} + \hat{a}_2 x_{i2} + \ldots + \hat{a}_k x_{ik} - y_i\right)^2.$$

Цель анализа - найти модель с наиболее точно предсказывающую информацию на имеющихся данных. Однако слишком фанатичная гонка за минимизацией ошибки может привести к **переобучению**, чтобы этого не случилось данные делят на **обучающие** (по которым строят модель и оценивают коэффициенты) и **тестовые**.

Среднеквадратичной ошибка помогает подчеркнуть большие ошибки и выбрать модель, которая дает меньше больших ошибок прогноза. Грубые ошибки становятся заметнее за счет того, что ошибку прогноза мы возводим в квадрат. И модель, которая дает нам

меньшее значение среднеквадратической ошибки, можно сказать, что что у этой модели меньше грубых ошибок.

Другой метод контроля качества модели - **коэффициент детерминации(R^2)**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Фактически, данная мера качества — это нормированная среднеквадратичная ошибка. Если она близка к единице, то модель хорошо объясняет данные, если же она близка к нулю, то прогнозы сопоставимы по качеству с константным предсказанием.

4.2 Простейшая модель линейной регресии

Построим модель линейной регресии для исследуеммых данных и сделаем выводы. Важно отметить, что хотя результат принимает значения только 1 или 0, предсказания модели не ограничены такими рамками. С учетом того, что известно, что 0 - согласие банка, 1 - отказ, к результатам можно относиться как к вероятности принадлежности к одному из результатов или же как к желанию банка одобрить заявку(чем ниже тем лучше)

Для построения модели использвуется библиотека sklearn и алгоритм:

1. Разделение выборки на обучающую и тестовую в соотношении 7:3
2. Инициализиция линейной регрессии
3. Тренировка модели на обучающих данных
4. Получение предсказания
5. Вывод полученной модели
6. Вывод оценки качества модели

В результате построения простейшей модели было получено :
свободный коэффициент = 1.57869864
коэффиценты линейной регрессии:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15
1	0.025843	-0.001898	0.005688	-0.542593	0.303663	-0.006254	-0.007839	-0.012319	-0.609144	-0.103397	-0.006875	0.023453	-0.000038	0.000051	-0.00002

Среднеквадратическую ошибка: 0.14
Коэффициент детерминации: 0.44
Это неплохие результаты, очевидно что полученное предсказание работает лучше чем случайное предсказание(под случайным предсказанием в этом случае можно подразумевать предсказание которое в среднем ошибается на 0.5) с таким предсказанием среднеквадратическую ошибка: 0.25
Интересно, что значение свободного коэффициент > 1, а значение большинства коэффицентов линейной регрессии отрицательны. Можно предполагать, что модель предсказания работает по принципу: изначально банк незаинтересован в одобрение кредита, но входные данные убеждает его рассмотреть этот вариант. Конечно на самом деле это не совсем так, ведь банк ориентируется на среднестатистического человека, а не на "нулевого"). Нулевой человек это, скорее всего, кто то вроде младенца, конечно, младенец - последний человек которому одобрят кредит.

4.3 Усложненная модель линейной регресии

Теперь построим усложненную модель линейной регресии.

Никто не гарантирует, что объясняемая переменная зависит от остальных характеристик именно линейно. Зависимость может быть, например, квадратичной или логарифмической; больше того, могут быть важны не только отдельные признаки, но и их комбинации.

Это можно учитывать, добавляя в качестве дополнительных признаков разные функции от уже имеющихся характеристик: их квадраты, логарифмы, попарные произведения. Разумеется, добавляя переменные стоит иметь в виду, что новая модель, не смотря на более больше колличество данных, может будет работать даже хуже исходной из за переобучения. Выбор переменных - сложная задача, у которой есть разные пути решения.

Попрубуем воспрользоваться информацией из разведочного анализа данных чтобы подобрать новые переменные, которые улучшат качество модели.

Обратимся к тепловой карте, показывающей корреляцию 2 признаков. Можно заметить, что есть есть умеренная корреляция в паре данных $A2$, $A8$

Попробуем также выбрать некоторое значение, которое изначально не было в формате чисел. Обратимся к графику с распределением значений столбцов с строками(буквами). Из всех столбцов у $A6$ наиболее широкий спектр значений

Таким образом, для усложнения модели будет использоваться $A2 \cdot A8$ и $(A6)^2$ квадрате

Далее добавим новые признаки в dataframe зависимых переменных и аналогично алгоритму построения простой модели, построим новую

В результате построения простейшей модели было получено : свободный коэффициент = 1.48312745
коэффиценты линейной регрессии:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A2_A8	A6^2
1	0.029735	-0.003132	0.005248	-0.504723	0.282833	0.030993	-0.003364	-0.026616	-0.582817	-0.103127	-0.007329	0.021556	-0.004625	0.000091	-0.00002	0.000341	-0.003033

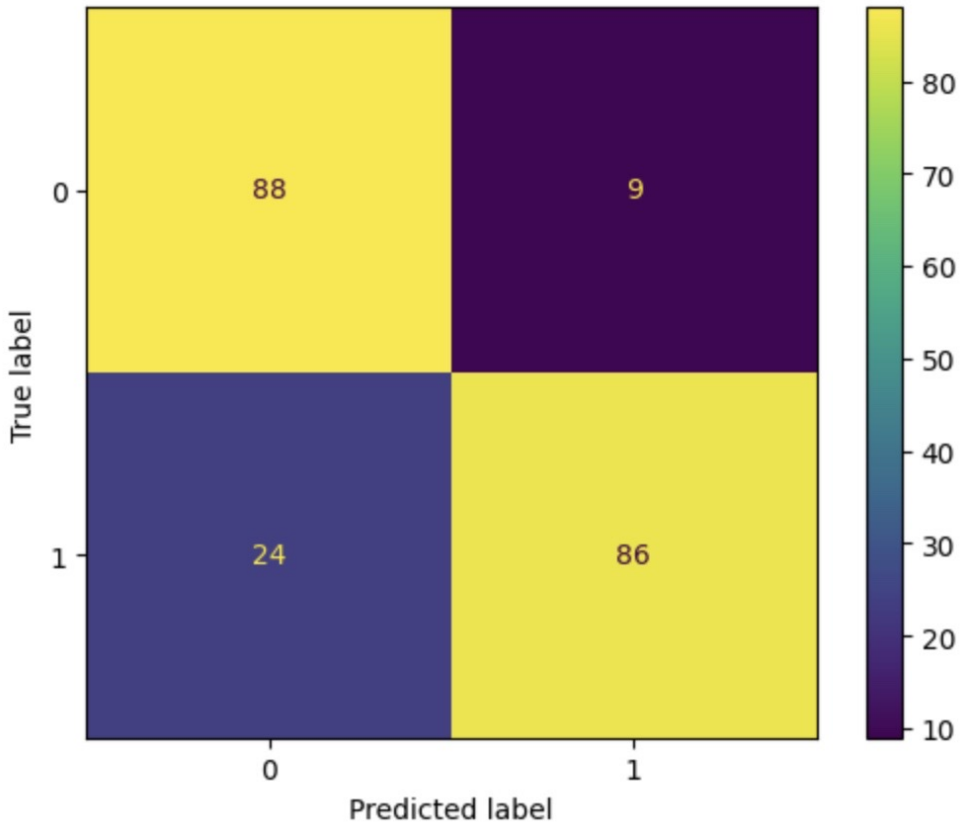
Среднеквадратическую ошибка: 0.13
Коэффициент детерминации: 0.46
среднеквадратическую ошибка уменьшилась, коэффициент детерминации увеличился, это показывает что усложненная модель более эффективна. Кроме того, что была найдена более хорошая модель, важно также отметить удачность выбранных признаков. Этот факт косвенно подтверждает, что есть зависимость между $A2$, $A8$ (не только числовая, но и логическая). Кроме того это показывает, что даже данные переведенные в численные значения действительно полезны(или хотя бы не вредны) при построении прогноза, иначе бы Среднеквадратическую ошибка увеличилась.

5 Модель классификации

Второй моделью было выбрана модель классификации, задачей которой является получение категориального ответа на основе набора признаков. Для решения этой задачи был выбран метод логистической регрессии. Логистическая регрессия - это алгоритм машинного обучения, который используется для решения задачи бинарной классификации, то есть разделения данных на два класса. Для построения модели использовалась библиотека **sklearn**. На первом шаге мы отмасштабировали данные - это необходимо для улучшения результатов анализа данных и обучения модели. Далее, так как данные уже

были разбиты на соответсвующие выборки(тренировочную и тестовую), модель была обучена на тренировочной выборке. После оставалась лишь оценить результаты работы на тестовой выборке. Для этого использовались методы оценки точности модели и матрица ошибок. Матрица ошибок - матрица размера 2 x 2, показывающая соотношение правильно и неправильно предсказанных положительных и отрицательных результатов. Пример матрицы:

Посмотрим на результаты нашей модели. Точность нашей модели оказалась 0.84, матрица ошибок выглядит следующим образом:



6 Заключение

В заключение, хотелось бы подвести итоге проведенной работы. Была проведена предварительная обработка данных и подготовка данных для построения моделей, построенно

две разные модели, понятных в использовании и показавших хорошую точность работы. Кроме построенных моделей, предсказывающих результат на основе зашифрованных данных, удалось также сделать несколько гипотез относительно логической составляющей, того как работает механизм рассмотрения решений в банках. Более того, удалось составить несколько предположений касательно зашифрованных данных, а именно о неслучайности их корреляции и наличие связи между столбцами, например между A2 и A8. Это данные впоследствие можно использовать для анализа данных,таких как построение других моделей или попытке расшифровки данных.