Займы

Цели проекта

- По финансовым характеристикам компании, а также по размеру и сроку займа предсказывать случится ли финансовый дефолт у компании.
- По средним финансовым характеристикам компаний бравших займ за месяц, а также по средним характеристикам взятых займов предсказывать процент дефолтов за тот же месяц.

Немного о данных

```
NUMERICAL_COLS = ["Term", "NoEmp",
                  "DisbursementYr", "DisbursementGross",
                  "CreateJob", "RetainedJob", ]
CATEGORICAL_COLS = ["DisbursementMo", "State", "NAICS"]
BINARY_COLS = [ "LowDoc", "RevLineCr", "Default"]
DATE_COLS = ["DisbursementDate"]
ALL_COLS = list(df.columns)
```

DisbursementGross - размер выданного займа
Term - Период на который взят займ в месяцах
NoEmp - Количество работников
CreateJob - Количество созданных рабочих мест
RetainedJob - Количество сохраненных рабочих мест
RevLineCr - Возобновляемый кредит
LowDoc - тип кредита, в котором требуется меньше документов
DisbursementDate - Дата выдачи займа
Default - Целевая переменная - Случился ли дефолт (0 выплатили, 1 - не выплатили займ)
Суммарный размер датасета - 899164 строк

State	NAICS	Term	NoEmp	CreateJob	RetainedJob	RevLineCr	LowDoc	DisbursementDate	DisbursementGross	Default
IN	45	84	4	0	0	0	1	1999-02-28	60000.0	0
IN	72	60	2	0	0	0	1	1997-05-31	40000.0	0
IN	62	180	7	0	0	0	0	1997-12-31	287000.0	0
OK	NaN	60	2	0	0	0	1	1997-06-30	35000.0	0
FL	NaN	240	14	7	7	0	0	1997-05-14	229000.0	0
...	...	...	...	...	...	...	...	...	...	...
TX	NaN	84	5	0	0	0	1	1997-06-30	79000.0	0
OH	45	60	6	0	0	1	0	1997-10-31	85000.0	0
CA	33	108	26	0	0	0	0	1997-09-30	300000.0	0
HI	NaN	60	6	0	0	0	1	1997-03-31	75000.0	1
HI	NaN	48	1	0	0	0	0	1997-05-31	30000.0	0

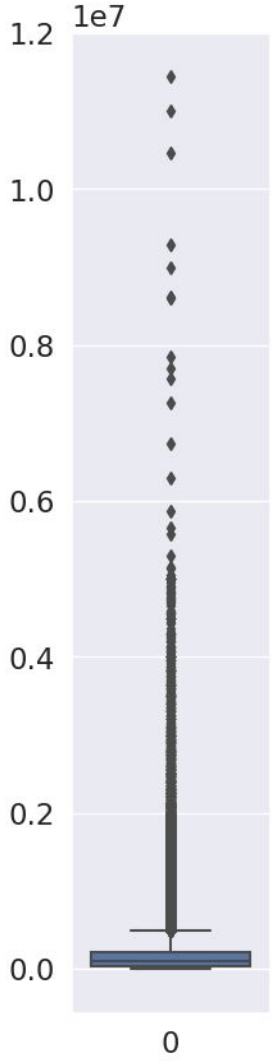
Про усредненный датасет

Также мы построили “усредненный датасет”, путем агрегации по всем месяцам каждого года:

Для численных переменных: агрегация средним - выборочное среднее значение в месяц.
Для бинарных переменных: агрегация средним - выборочный процент положительных исходов бинарной переменной в месяц.

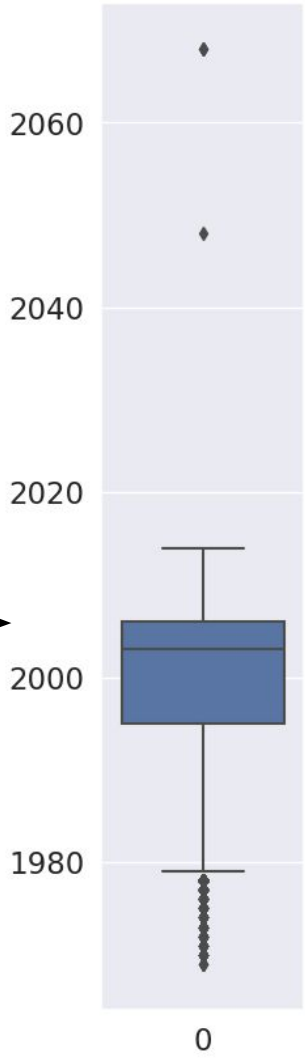
	DisbursementDate	Средний Term	Средний NoEmp	DisbursementYr	Средний DisbursementGross	Средний CreateJob	Средний RetainedJob	Процент NewBusiness	Процент LowDoc	Процент RevLineCr	Процент Default
237	1990-01	131.789110	21.744413	1990.0	135441.937500	0.619260	0.469321	0.275498	0.0	0.000000	0.058106
238	1990-02	183.552846	19.268293	1990.0	166335.312500	7.195122	7.756098	0.317073	0.0	0.000000	0.056911
239	1990-03	177.253425	19.082192	1990.0	154888.468750	5.856164	5.938356	0.294521	0.0	0.000000	0.041096
240	1990-04	128.984592	16.271327	1990.0	136543.718750	0.417888	0.290492	0.264186	0.0	0.000000	0.046223
241	1990-05	147.299094	13.187311	1990.0	144547.390625	2.945619	2.848943	0.247734	0.0	0.000000	0.066465
...	...	...	...	...	...	...	...	...	...	...	...
522	2013-10	72.830769	11.692308	2013.0	92126.320312	3.369231	7.492308	0.230769	0.0	0.461538	0.046154
523	2013-11	73.594595	5.405405	2013.0	75329.093750	3.040541	4.418919	0.337838	0.0	0.486486	0.040541
524	2013-12	63.585714	9.071429	2013.0	93398.414062	2.457143	5.614286	0.257143	0.0	0.571429	0.042857
525	2014-01	63.787879	14.909091	2014.0	81725.625000	3.757576	7.818182	0.257576	0.0	0.484848	0.060606
526	2014-02	63.250000	7.325000	2014.0	56369.800781	2.425000	1.775000	0.300000	0.0	0.525000	0.025000

Первичное удаление выбросов

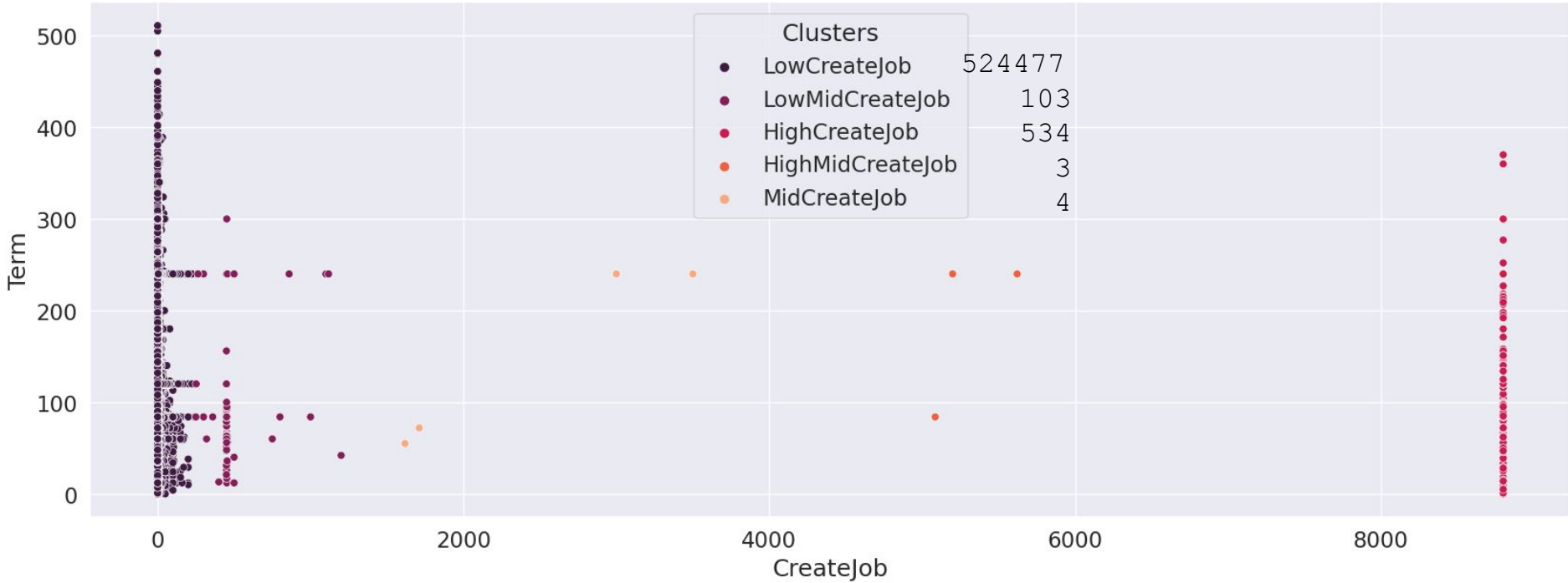


DisbursementGross

DisbursementYr

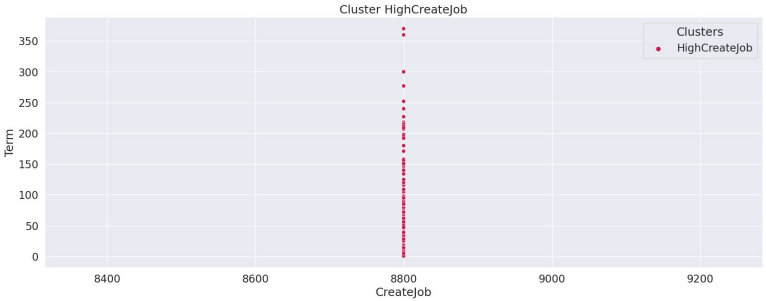


Кластеризация по CreateJob (KMeans)

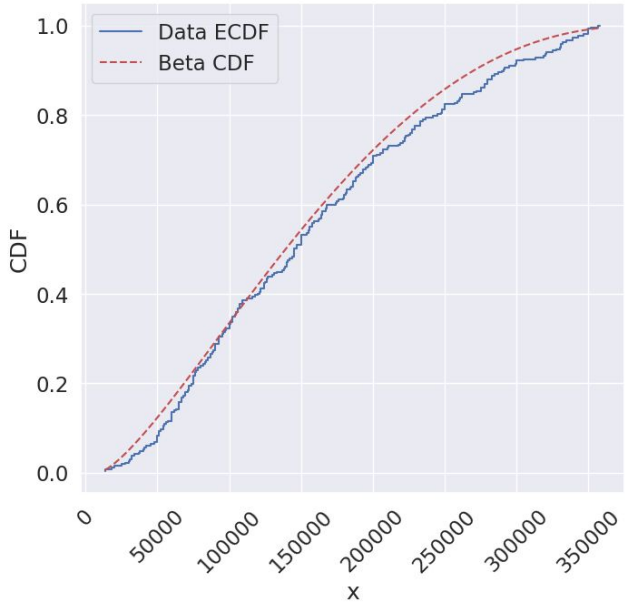


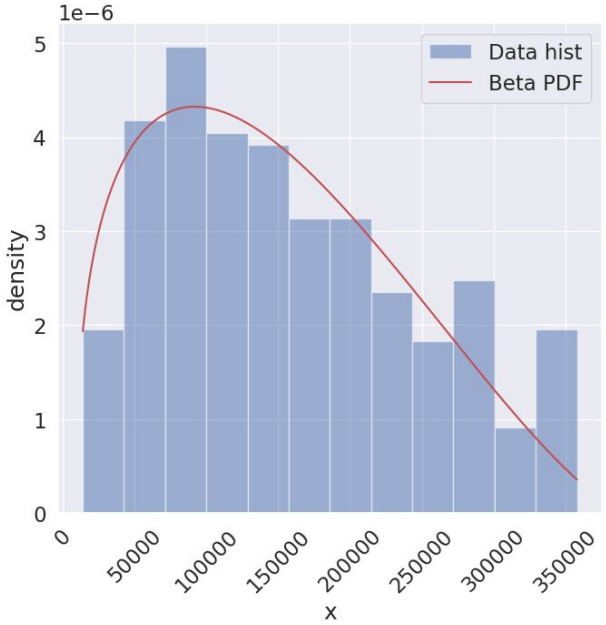
Kolmogorov-Smirnov test для DisbursementGross для крупных кластеров

- Для проведения теста часть выборки, которая участвует в тесте случайным образом делилась пополам. Для первой части получали оценки максимального правдоподобия. Для второй части, используя найденные ранее оценки проводили kstest, а также строили гистограммы и функции распределения.

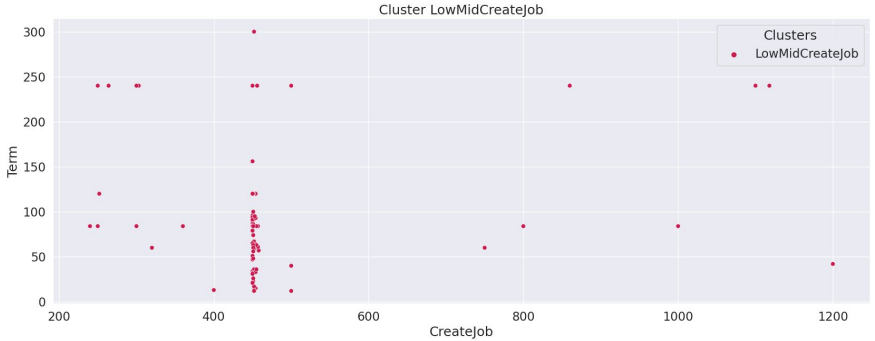


DisbursementGross: Cluster HighCreateJob
KstestResult(statistic=0.06, pvalue=0.28614138261217226)

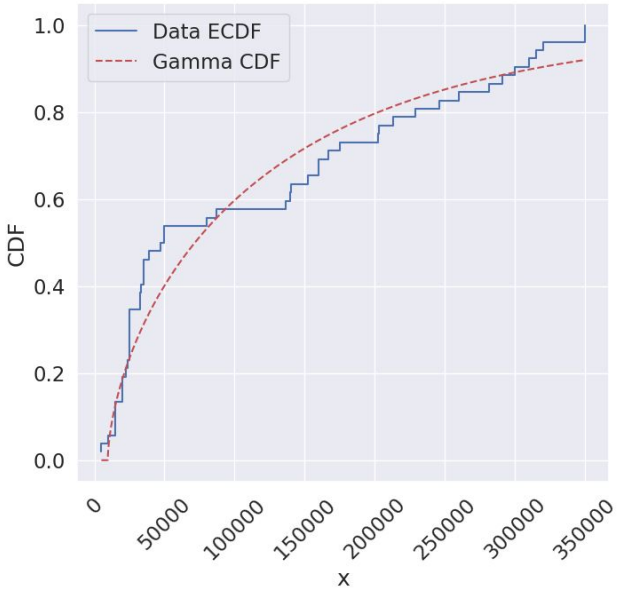


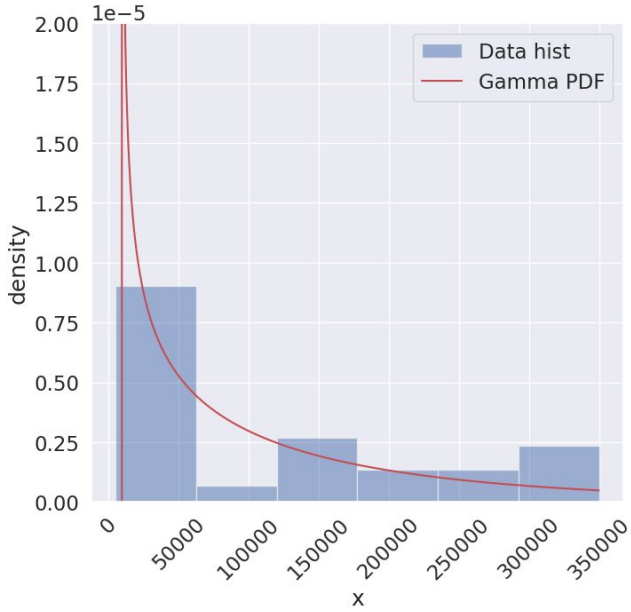


- Для кластера, где CreateJob высокий, нельзя отвергнуть гипотезу о том, что DisbursementGross имеет beta распределение.



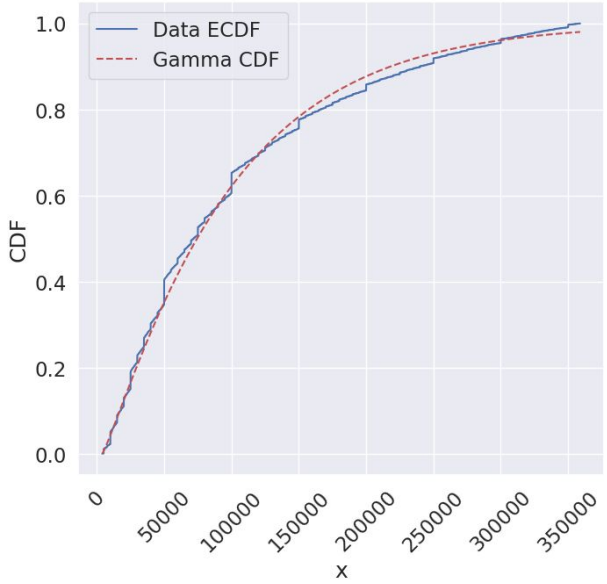
DisbursementGross: Cluster LowMidCreateJob
KstestResult(statistic=0.13, pvalue=0.2963522271243325)

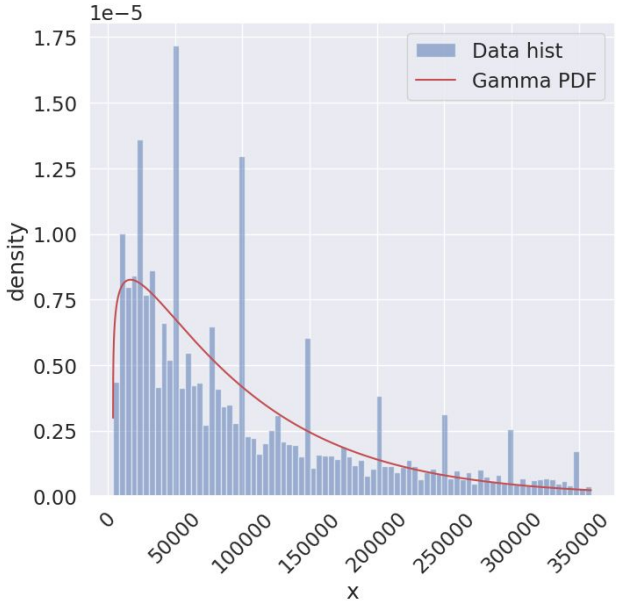




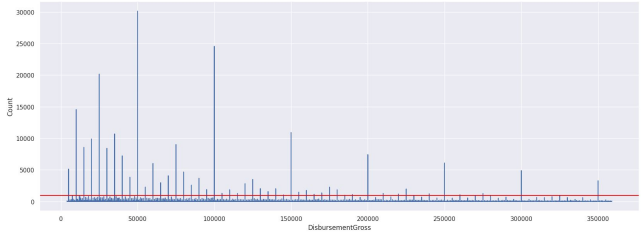
- Для кластера, где CreateJob чуть выше низкого, нельзя отвергнуть гипотезу о том, что DisbursementGross имеет gamma распределение.

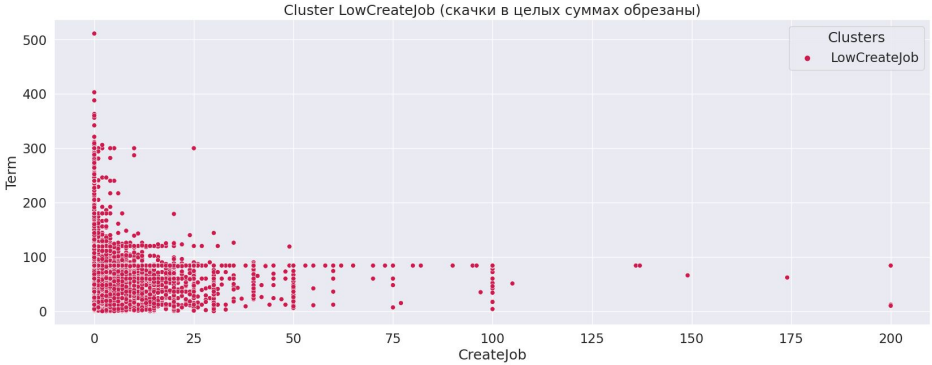
DisbursementGross: Cluster LowCreateJob (скачки в целых суммах не обрезаны)
KstestResult(statistic=0.05, pvalue=0.0)



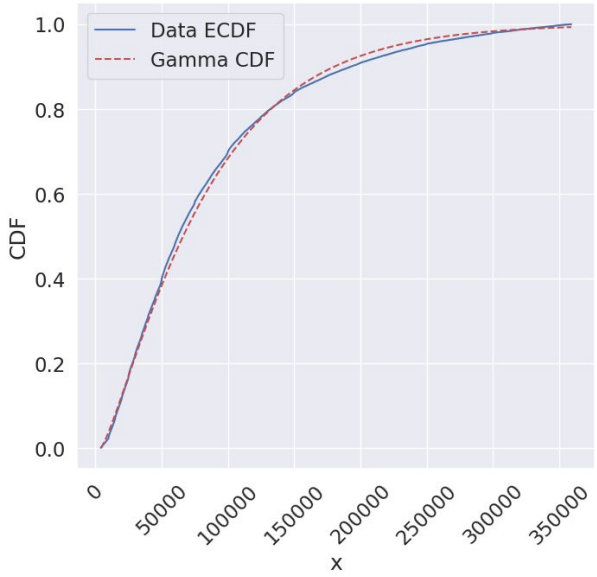


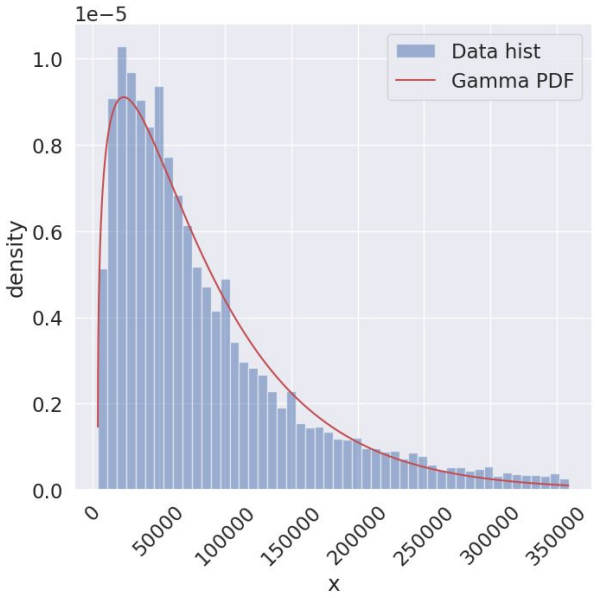
- Для кластера, где CreateJob низкий (самый большой) распределение DisbursementGross неоднозначно.
- Минимальное значение статистики достигается для гипотезы о Gamma распредении.
- Видно, что скачки в “круглых” суммах портят картину.





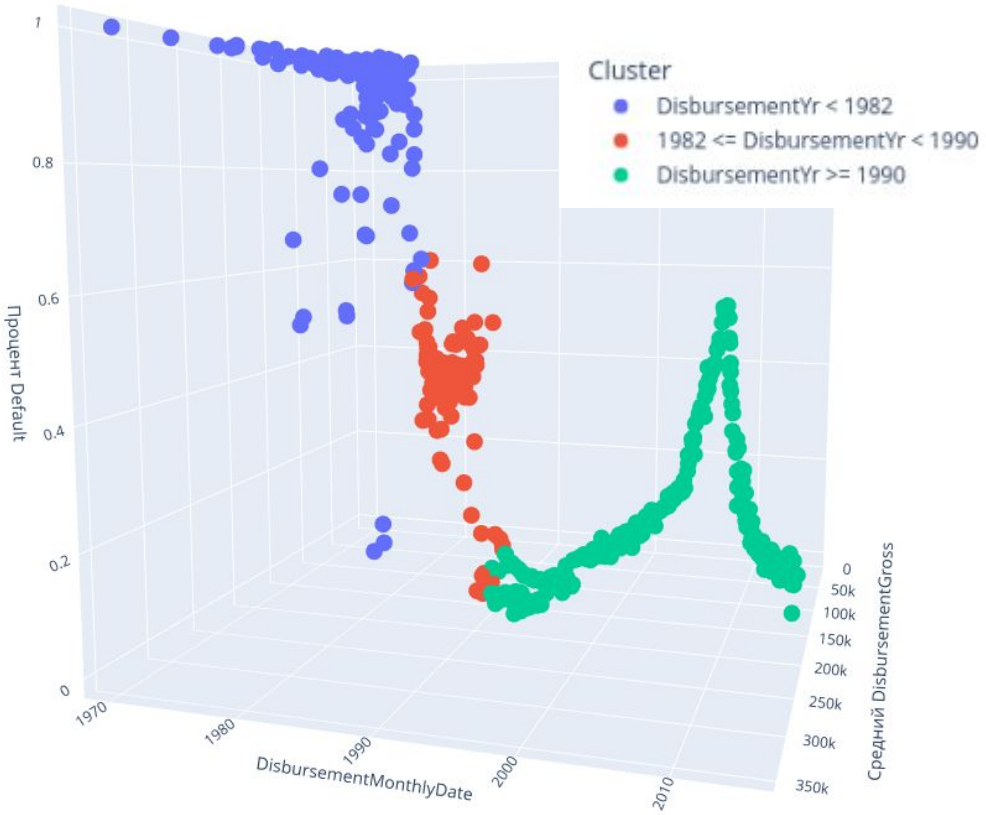
DisbursementGross: Cluster LowCreateJob (скачки в целых суммах обрезаны)
KstestResult(statistic=0.03, pvalue=1.4326852639814262e-42)

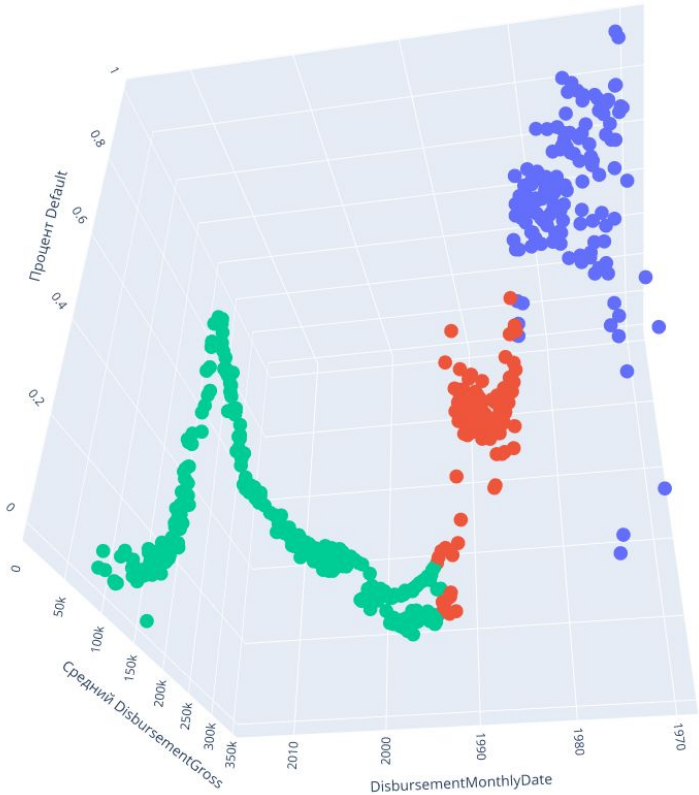




- После удаления дискретной составляющей, гипотеза о том, что DisbursementGross имеет gamma распределение имеет смысл. p-value получился ненулевой.
- Учитывая специфику рассматриваемых данных (большое кол-во данных, а также остаточные скачки) можно заключить, что основную гипотезу сложно отвергнуть.

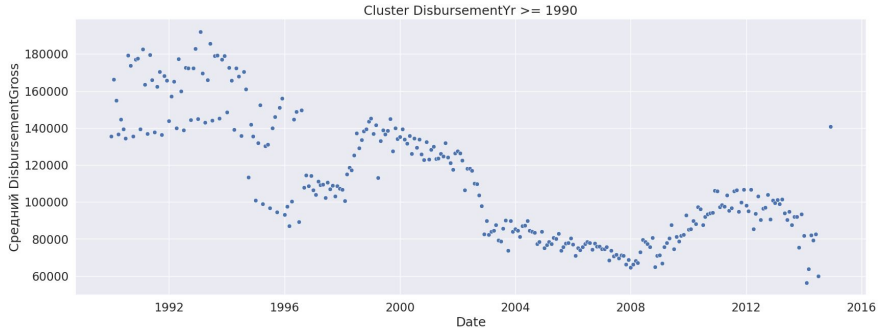
Кластеризация по годам в усредненном датасете



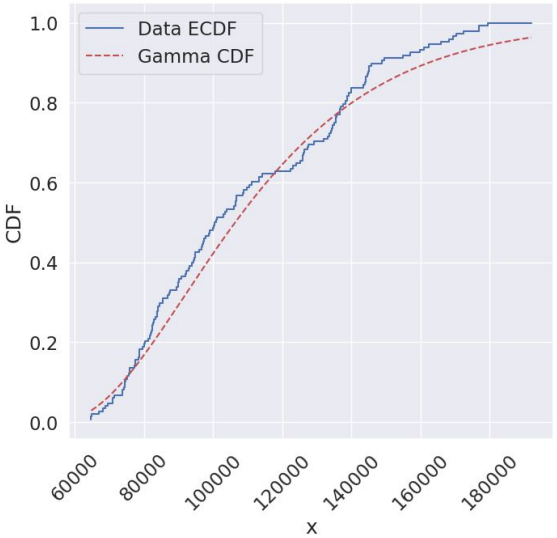


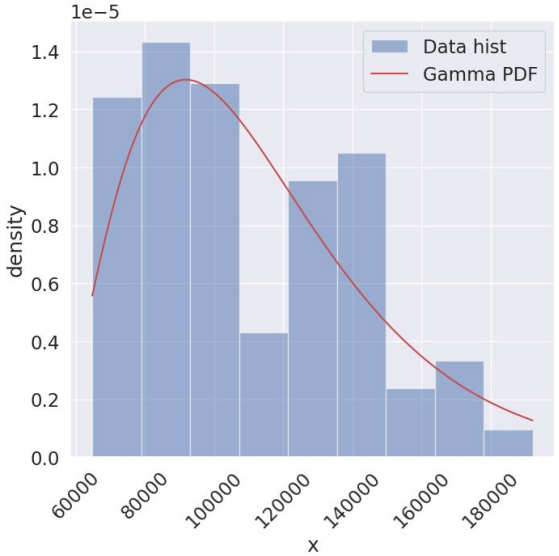
Распределение данных по выбранным кластерам

Kolmogorov-Smirnov test для среднего DisbursementGross для самого крупного кластера (DisbursementYr >= 1990)



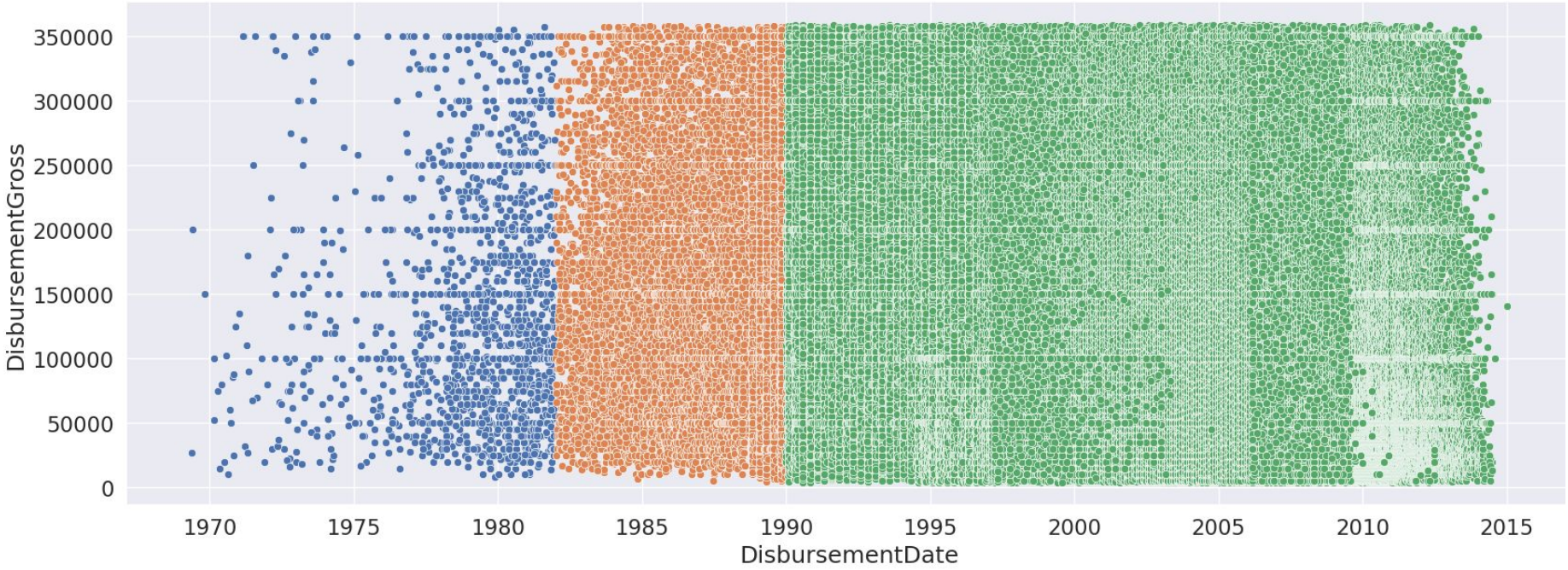
Средний DisbursementGross: Cluster DisbursementYr >= 1990
KstestResult(statistic=0.07, pvalue=0.43842365231563585)



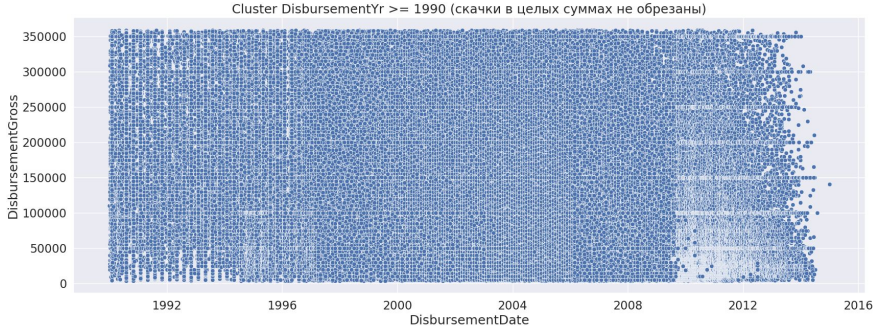


- Для данного кластера нельзя отвергнуть гипотезу о том, что Средний DisbursementGross имеет gamma распределение.
- Кроме того, стоит отметить, что данный кластер - самый большой и современный.

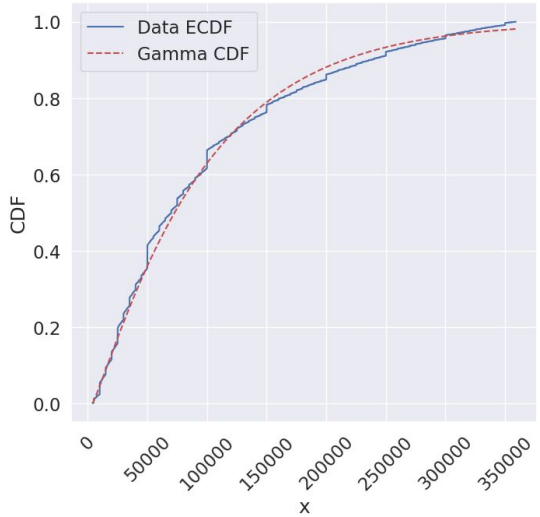
Перенесем кластеризацию по годам на оригинальный датасет. Посмотрим опять же на самый большой кластер (DisbursementYr >= 1990).

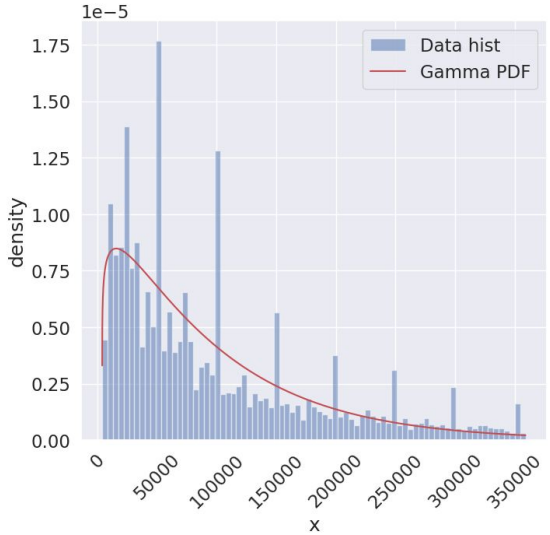


Kolmogorov-Smirnov test для DisbursementGross для самого крупного кластера (DisbursementYr >= 1990)



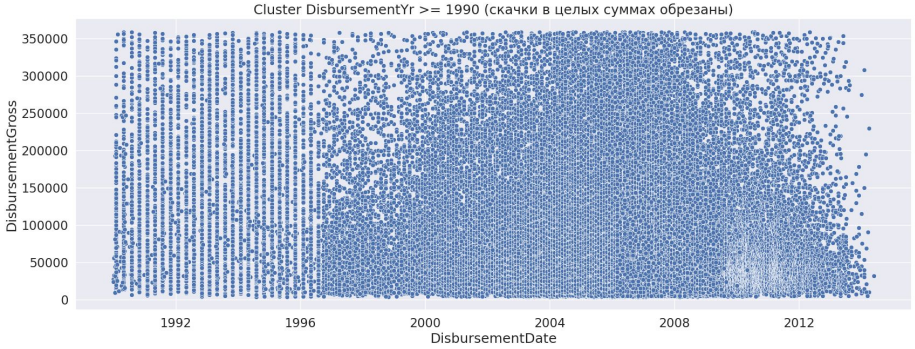
DisbursementGross: Cluster DisbursementYr >= 1990 (скачки в целых суммах не обрезаны)
KstestResult(statistic=0.05, pvalue=0.0)



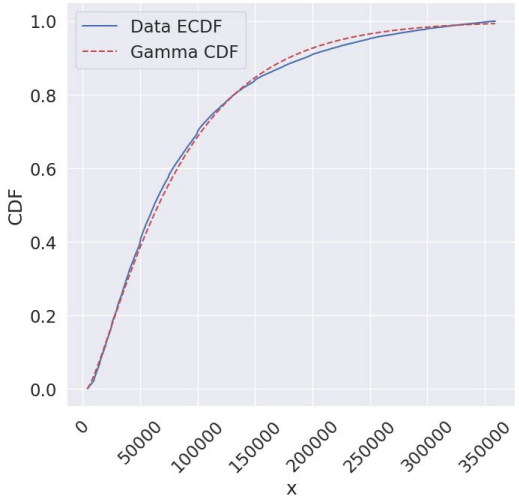


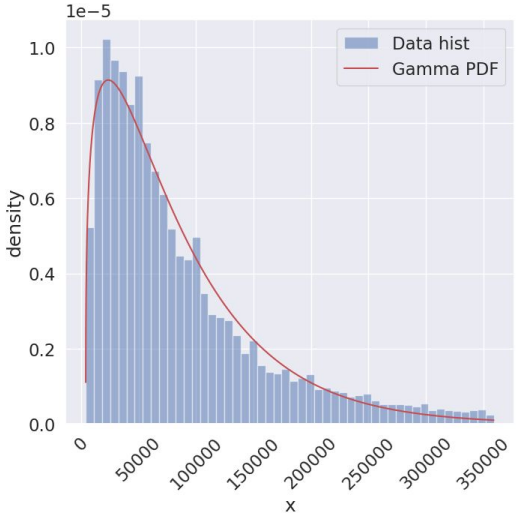
- Опять распределение DisbursementGross неоднозначно из-за скачков в “круглых” суммах.

Kolmogorov-Smirnov test для DisbursementGross для самого крупного кластера (DisbursementYr >= 1990)



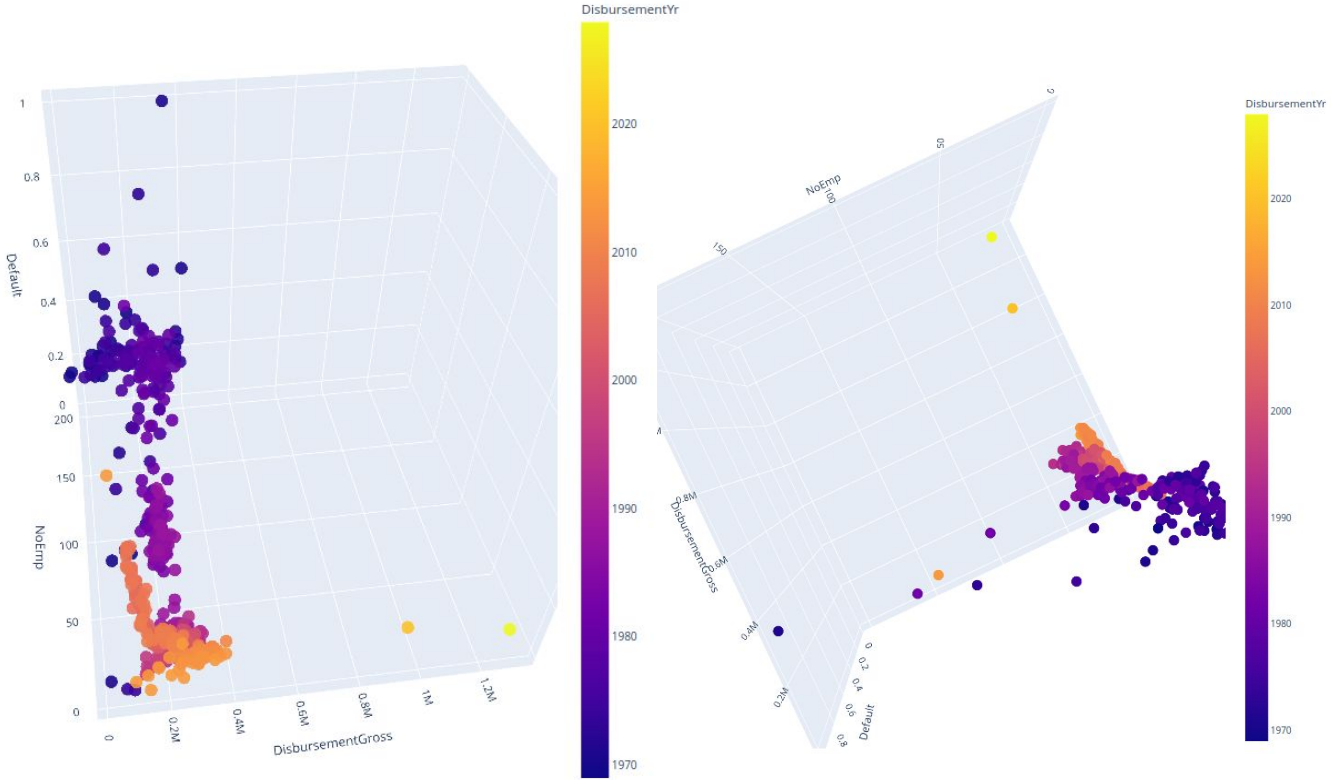
DisbursementGross: Cluster DisbursementYr >= 1990 (скачки в целых суммах обрезаны)
KstestResult(statistic=0.03, pvalue=1.4165093952861612e-38)





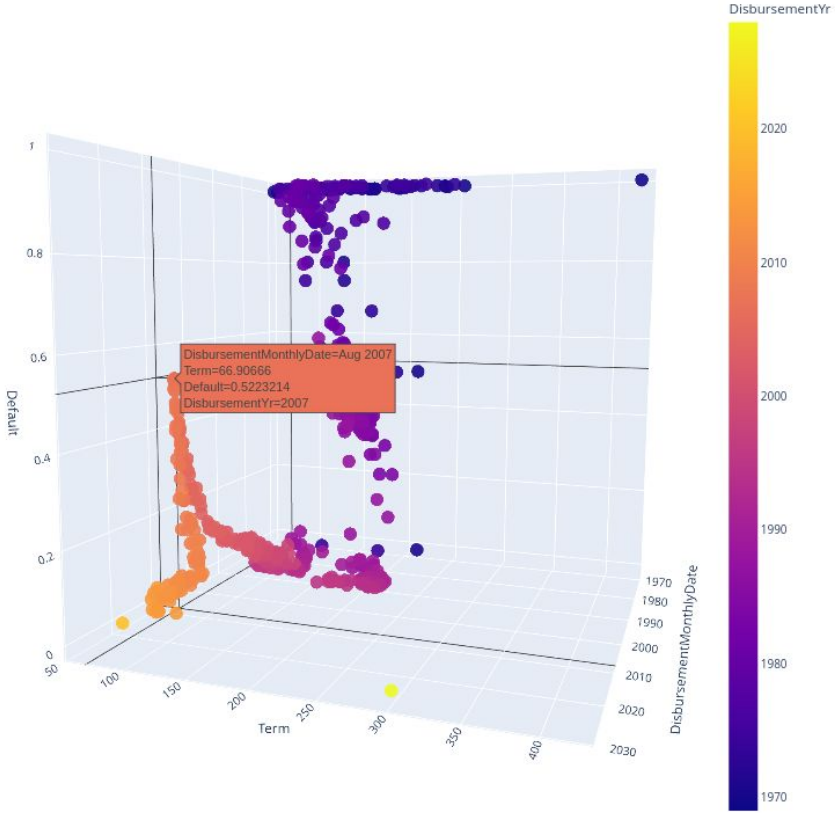
- После удаления дискретной составляющей, гипотеза о том, что DisbursementGross имеет gamma распределение вновь имеет смысл. Опять считаем, что ее сложно отвергнуть.

Связь средних NoEmp, DisbursementGross, Default %



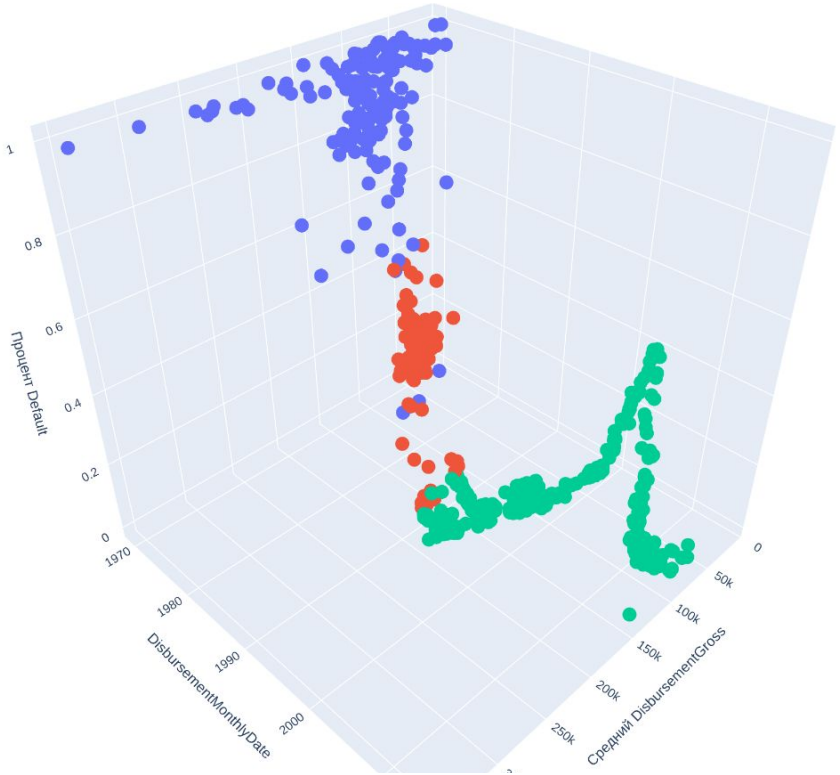
Если убрать старые данные (< 1985 г.) - более маленький и менее актуальный кластер данных, то прослеживается понижение процентов дефолтов с одновременным ростом размера займа и числа рабочих.

Средний Term vs Date vs Default %



1) Чем ближе к настоящему, тем меньше срок займа.
2) Пик соответствует фин.кризису.
3) Если отбросить старые данные (< 1985), то прослеживается снижение процента дефолтов с ростом срока займа, но после кризиса что-то изменилось.

Чистим данные перед регрессиями



По итогам анализа кластеров:

1. Наиболее репрезентативными являются данные в период с 1990 по 2015 (зеленые на графике). Удаляем все данные до 1990.
2. Удаляем данные, которые не принадлежат самому большому кластеру “LowCreateJob”

Строим логистическую регрессию для “не усредненного” датасета.

Предобработка данных:

- Балансировка датасета: Undersampling
  (строк, где Default = 0 сильно больше, чем Default = 1, поэтому берем sample от Default = 0 такой же по размеру, как и Default = 0)
- Масштабирование датасета: data -> StandardScaler

Ключевые Метрики:

- F1
- Recall
- Precision
- Accuracy

Дополнительно:

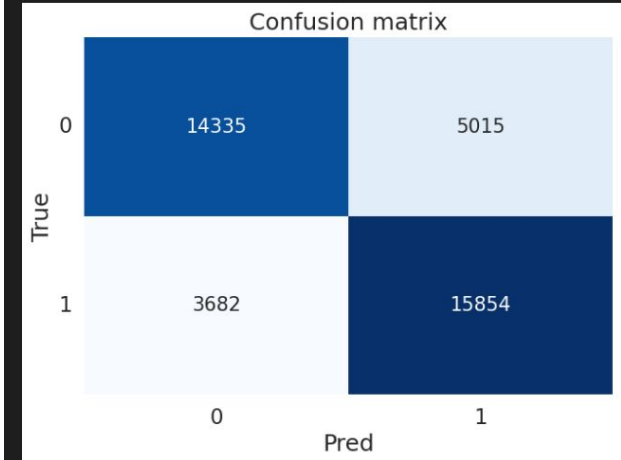
- PR curve
- ROC curve

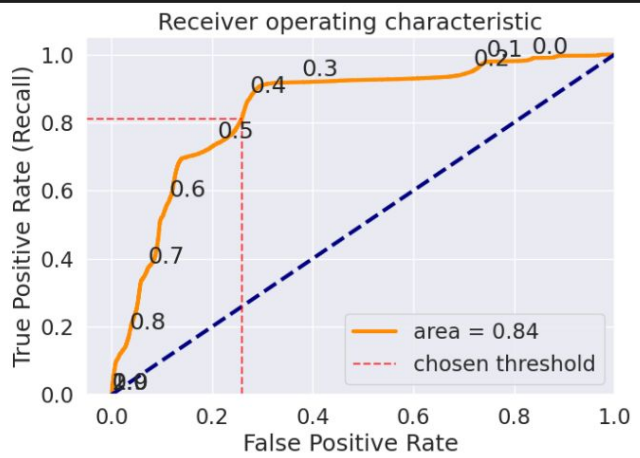
Метрики

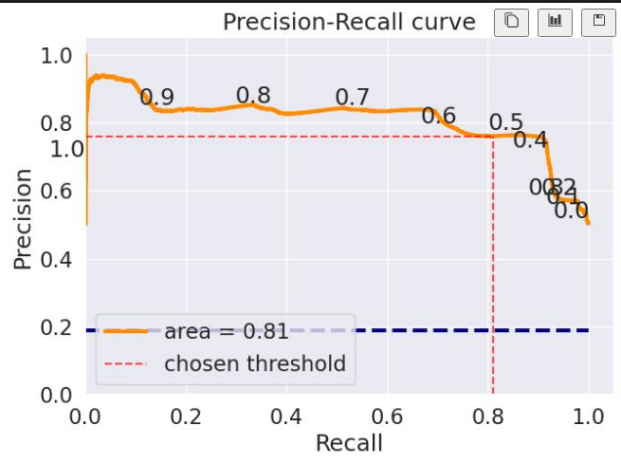
```
Accuracy (overall correct predictions): 0.78
Auc: 0.84
Recall (all 1s predicted right): 0.82
Precision (confidence when predicting a 1): 0.75
F1 score: 0.78
```

Коэффициенты

```
DisbursementGross: Weight -0.02437044243021235
NoEmp: Weight -0.9079613663845036
Term: Weight -2.1333150942100794
CreateJob: Weight 0.17844048364626094
RevLineCr: Weight -0.053438784105272905
```







Делаем признаки полиномиальными

Признаки оставляем такими же, но делаем их полиномиальными (степень=2) и запускаем на них лог. регрессию

Пример: если у нас есть набор признаков: x1, x2, x3. То из них мы делаем x1, x2, x3, x1^2, x2^2, x3^2, x1*x2, x2*x3, x3*x1

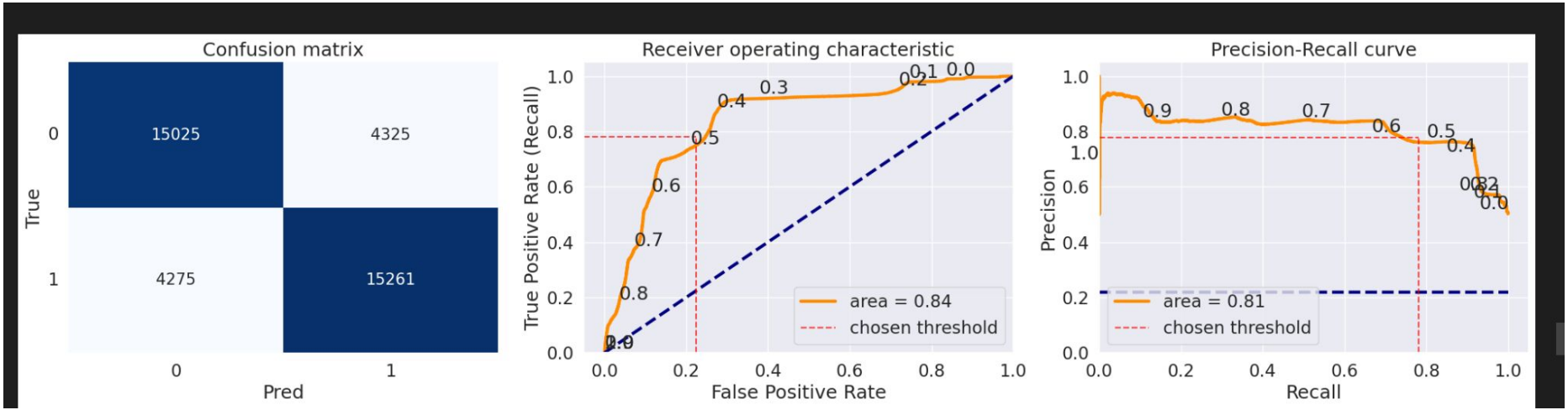
Метрики такие же

Получившиеся коэффициенты:

```
DisbursementGross: Weight -0.575245829860737
NoEmp: Weight -2.1048597006533556
Term: Weight -2.4038156776783324
CreateJob: Weight 1.0635708994081554
RevLineCr: Weight 0.42237289971910646
DisbursementGross^2: Weight 0.12513082904332287
DisbursementGross NoEmp: Weight 0.612906773965541
DisbursementGross Term: Weight 0.9128456363278872
DisbursementGross CreateJob: Weight -0.3408283915104417
DisbursementGross RevLineCr: Weight 0.023997613162685495
NoEmp^2: Weight 1.2429474618431857
NoEmp Term: Weight -0.26155451573059335
NoEmp CreateJob: Weight -0.032363969323877015
NoEmp RevLineCr: Weight 0.3590296168348376
Term^2: Weight 0.625227878201182
Term CreateJob: Weight -0.49581763512663574
Term RevLineCr: Weight -0.9086279678264422
CreateJob^2: Weight -0.2435760227426206
CreateJob RevLineCr: Weight -0.3215222200522381
RevLineCr^2: Weight 0.4223728997190777
```

Метрики

```
Accuracy (overall correct predictions): 0.78
Auc: 0.86
Recall (all 1s predicted right): 0.79
Precision (confidence when predicting a 1): 0.77
F1 score: 0.78
```



Линейная регрессия на усредненном датасете

Цель: определить какой процент займов закончится дефолтом в текущем месяце
Пояснение 1: У нас есть финансовые по всем займам в этом месяце, какова вероятность, что “процент Default” будет равен 1, чем он выше, тем соответственно хуже
Предобработка данных:

1. Применяем logit к данным (log(y/(1-y)), где y - целевая переменная
2. Приводим данные к одному масштабу

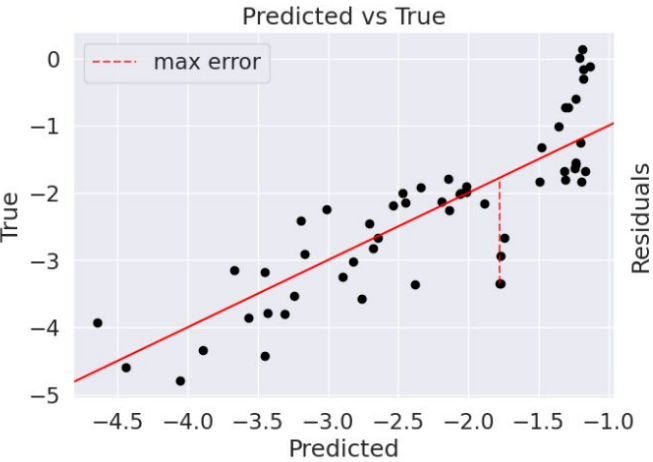
Пояснение 2: После получения prediction на test данных мы можем вернуться к исходным переменным, применив обратную к logit функцию, а можем остаться в текущем

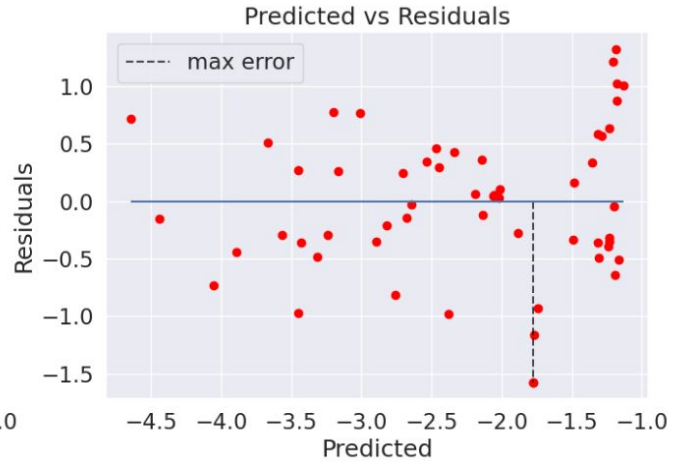
Ключевые Метрики:

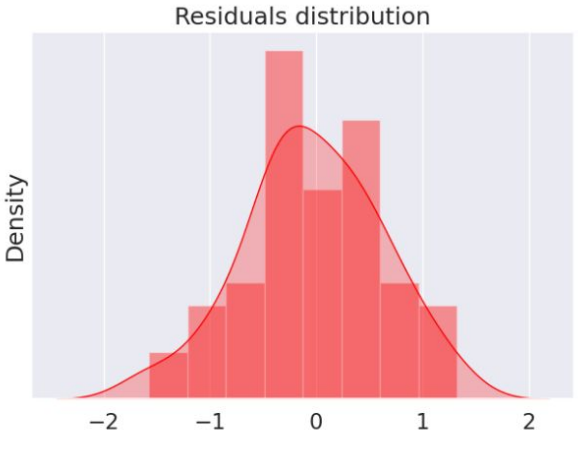
- R2
- MAPE
- Max Error

Результаты без применения функции обратной к logit

```
R2 (explained variance): 0.71
Mean Absolute Perc Error (Σ(|y-pred|/y)/n): 0.29
Mean Absolute Error (Σ|y-pred|/n): 0.515
Root Mean Squared Error (sqrt(Σ(y-pred)^2/n)): 0.642
Max Error: -2
```

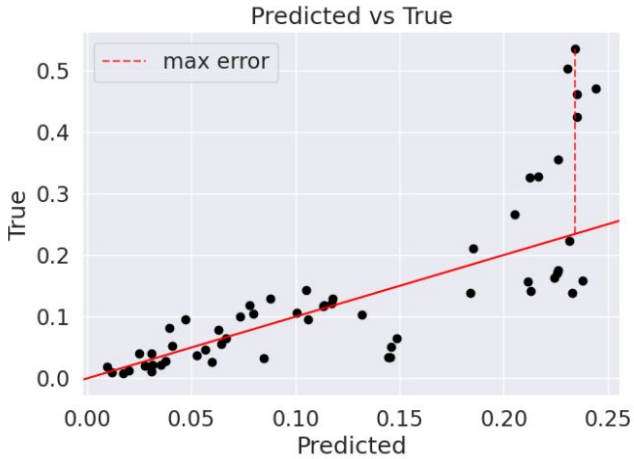


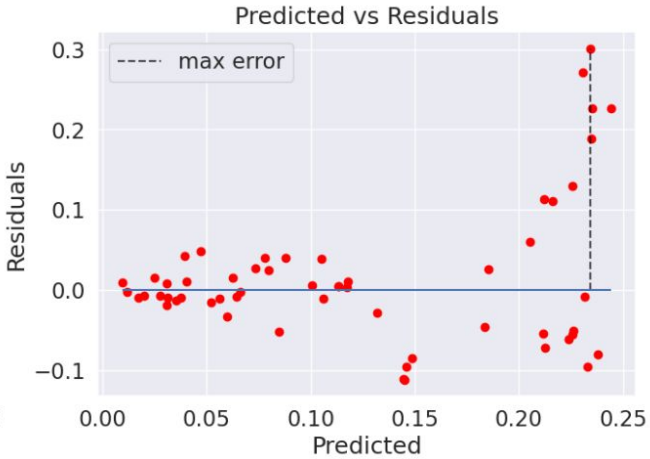


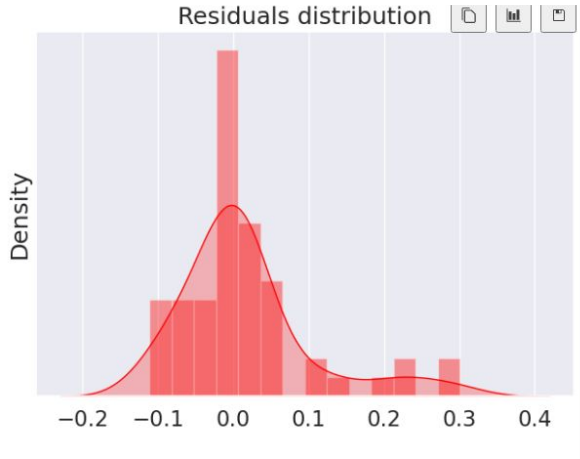


Результаты с применением функции обратной к logit

```
R2 (explained variance): 0.56
Mean Absolute Perc Error (Σ(|y-pred|/y)/n): 0.43
Mean Absolute Error (Σ|y-pred|/n): 0.057
Root Mean Squared Error (sqrt(Σ(y-pred)^2/n)): 0.089
Max Error: 0
```

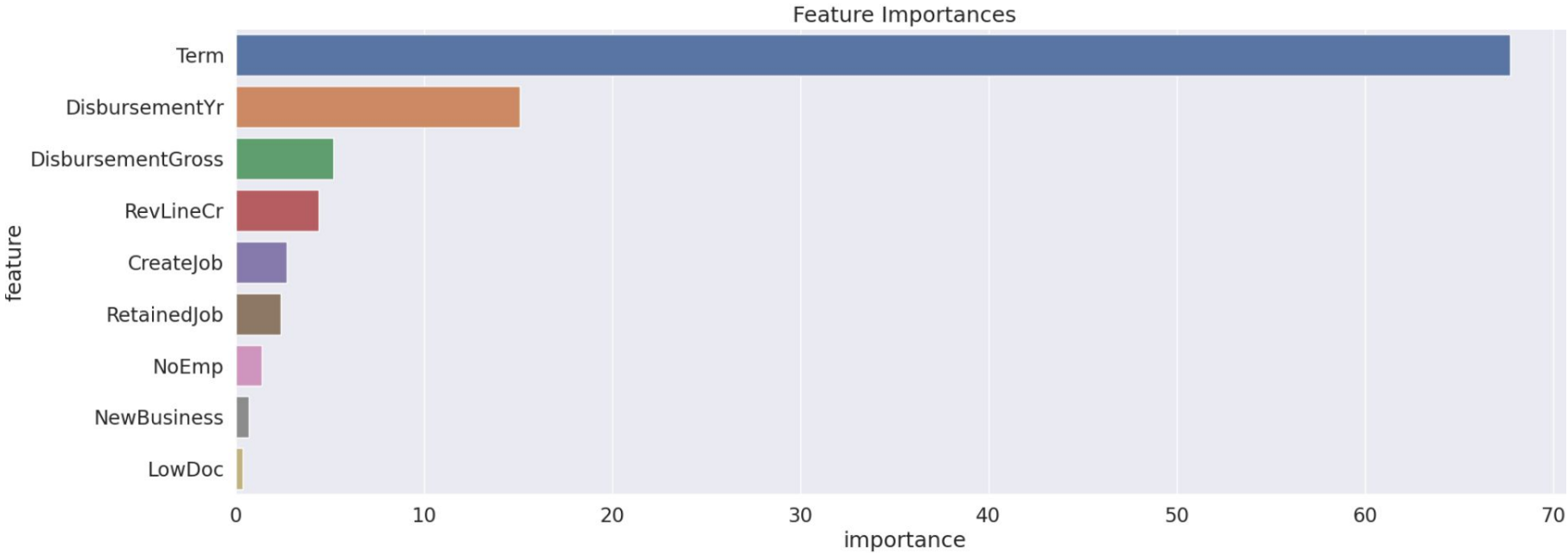






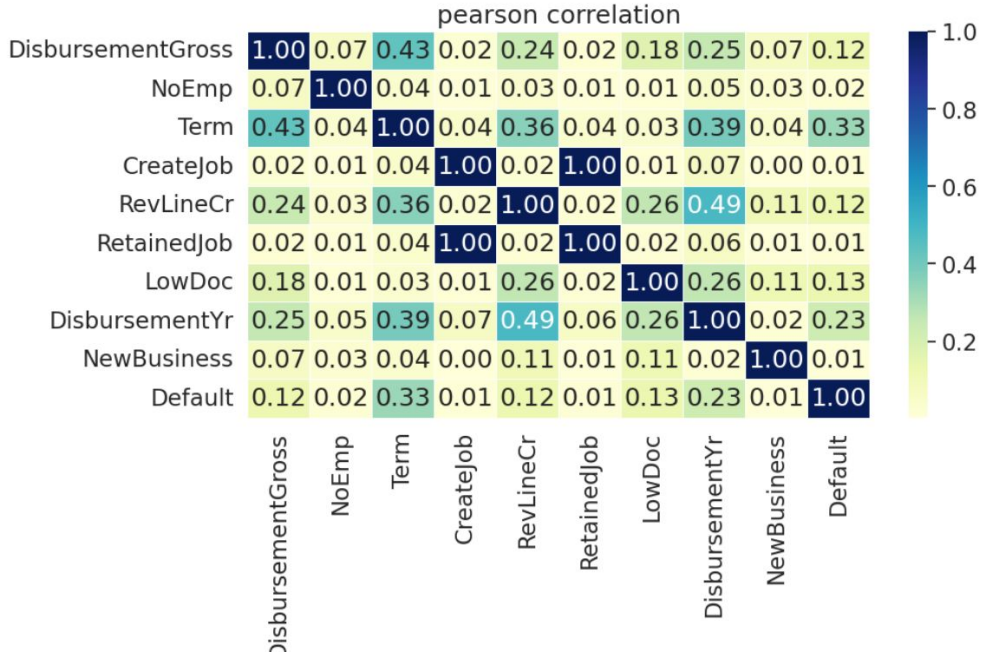
Проверка себя с помощью catboost feature importances

Строим модель CatBoostClassifier, используя все числовые признаки. Применяем feature importances

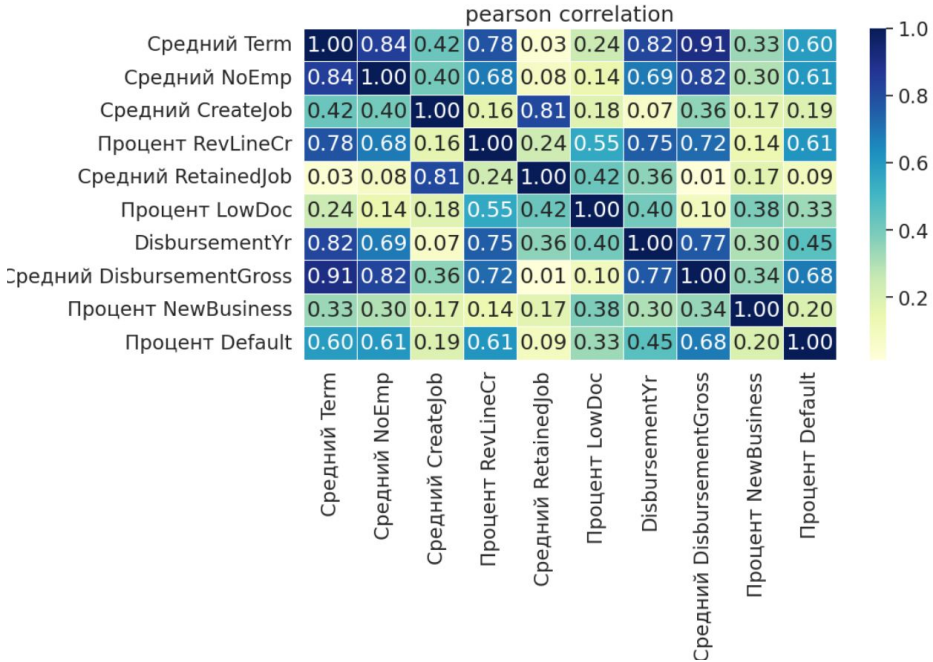


Матрицы корреляций

Соотносится с результатами catboost



Матрица корреляций для усредненного датасета



Глобальный вывод

Вероятность дефолта в первую очередь определяется запросом - (размер + срок займа) и внешними факторами - (год), а только потом уже финансовыми характеристиками компании