



Команда  
F.R.I.E.N.D.S.

# Прогнозирование просрочки по контрагенту

Евгений Васильев  
Игорь Ситник

# Постановка задачи

- ✗ На основе имеющихся данных мы будем обучать модель, которая будет предсказывать ПДЗ (столбец среднее ПДЗ) для компании на основе ее различных финансовых показателей



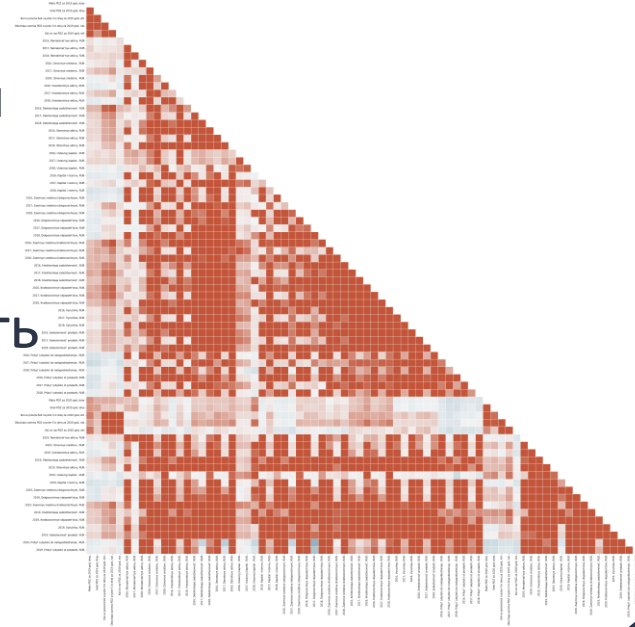
# Предобработка / чистка данных

- ✗ Склеиваем таблицы
- ✗ Удаляем компании, у которых ПДЗ больше года
- ✗ Избавляемся от значений с выбросами - в столбцах с финансами заменяем все выбросы на  $3\sigma$
- ✗ Нормируем столбцы с финансами (z-score)



# Корреляция между столбцами

X Максимальная корреляция 0.3, высокоскоррелированные столбцы выделять и убирать не нужно



# Модели машинного обучения



Scipy LinearRegression

- + Высокая «понятность»: вес каждого столбца в результат
- Низкая точность решения



Градиентный бустинг деревьев LightGBM

- + Высокая точность работы (5x по сравнению с LR)
- Низкая «понятность» решения (сложность интерпретации всех деревьев)
- Проблемы с запуском LightGBM на компьютерах Apple



# Качество предсказания

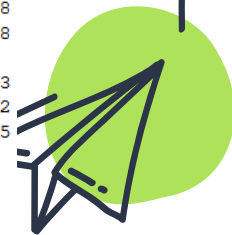
X LightGBM:

X RMSE of train prediction is: **6.87 days of PDZ**

X RMSE of test prediction is: **6.87 days of PDZ**

Real PDZ vs Predict PDZ

```
[(0.0, 0.3184776679155854), (2.230769230769231, 2.2313607913457902), (0.0, 0.14882691135021892), (2.0, 2.0379810872603703), (1.0, 1.68975210659412), (3.6875, 4.752579287360798), (1.285714285714286, 5.517705301501126), (0.0, 0.3093737575160942), (0.0, 0.1608612973047478), (0.0, 0.2558285608599299), (0.0, 0.37129420815916064), (4.028409090909091, 4.812839394295932), (13.05555555555556, 44.52165918620892), (4.125, 5.177495597631054), (0.0, 0.6865030957275444), (4.888888888888889, 5.061362981853052), (0.0, 0.31703596310177695), (0.0, 0.026266389246912412), (3.886178861788618, 2.795713062900803), (0.0, 1.2851609741393555), (1.0, 1.1175475456655046), (0.0, 0.5564979519464874), (0.0, 2.5154458045179178), (68.09090909090909, 49.60227184034706), (5.0, 4.752064854750671), (1.0, 1.7092949100119292), (15.5, 11.816802870835382), (7.422222222222222, 4.525680094166303), (12.3125, 6.382823828546713), (11.0, 10.874289922239333), (0.0, 0.47848727586877554), (5.0, 4.367903958791516), (5.0, 4.129358557517205), (17.66666666666667, 13.482654066373104), (0.0, 0.32875147830688106), (0.0, 2.558524869664252), (0.0, 0.25915789178519827), (0.0, 0.37534650636908484), (12.3333333333333, 12.89781024875428), (0.0, 0.371602662680674), (6.2, 5.344424665570279), (0.0, 1.105143225414187), (7.0, 11.79234900959343), (0.0, 2.0067375488340162), (6.25, 9.163656610070012), (0.0, -0.20261494972067534), (8.125, 8.667048433542206), (9.0, 10.424274342320217), (4.522388059701493, 6.492257006350921), (0.0, 0.37100666825018563)]
```



# Внедрение новых признаков для данных

Х коэффициент ликвидности абсолютной:

$$X \text{ КЛабс} = \frac{\text{ДенСр} + \text{КрФинВл}}{\text{КрКр} + \text{КрКредЗад} + \text{ПрОбяз}}$$

Х коэффициент ликвидности срочной:

$$X \text{ КЛср} = \frac{\text{ДенСр} + \text{КрФинВл} + \text{КрДебЗад}}{\text{КрКред} + \text{КрКредЗад} + \text{ПрОбяз}}$$



# Качество предсказания с новыми признаками

X LightGBM:

X RMSE of train prediction is: **6.97 days of PDZ**

X RMSE of test prediction is: **6.75 (0.12 better) days of PDZ**

Real PDZ vs Predict PDZ

```
[(5.26, 5.266798495106416), (2.25, 2.130370950508777), (0.0, 0.09766996041400978), (5.0, 3.8976785177454953), (2.4545
45454545455, 6.001245038111608), (5.739130434782608, 12.34322584021027), (0.0, 0.27781536293793696), (0.0, 0.48970465
94220893), (1.238095238095238, 1.6112510116365288), (3.1, 3.4406758013754195), (3.0, 3.042668078338976), (0.0, 0.0411
3280451890544), (19.0, 42.35711350007193), (0.0, 1.4978140139703273), (17.84210526315789, 36.18051194425227), (1.5,
1.921057409448517), (0.0, -0.17321463904435902), (9.666666666666666, 10.273342641663351), (0.0, 1.4246927416837807),
(0.0, 1.4910416647388252), (3.5, 4.3018519865629266), (0.0, 0.32408317121543984), (5.2, 9.86462750832846), (0.0, 0.54
92323988391139), (2.888888888888889, 5.3739053714636), (0.0, 0.4393396827855795), (3.725, 6.5754008872633225), (13.01
369863013699, 34.34523754863687), (2.428571428571428, 4.024494102400365), (5.576923076923077, 6.703791689282296), (0.
0, 0.42172638231446996), (3.059523809523809, 3.3636331020857457), (1.858870967741935, 2.2528824042589988), (79.104278
39958699, 42.06254415528244), (0.0, 0.19318656362536374), (0.0, -0.7714663928714989), (1.0, 1.031163776792429), (13.
0, 12.108407115701084), (3.0, 7.357810924467498), (2.348837209302326, 2.7661239222417104), (0.0, 1.5756311171427415),
(8.05586592178771, 12.500693861651838), (1.0, 1.348754833157988), (4.0, 3.6675537610648816), (5.251101321585903, 4.22
6013137564795), (3.0, 4.15745144201805), (0.0, -0.06685509986428534), (7.0, 5.668753418843815), (4.0, 4.7627178252827
75), (0.0, 0.8131512517068853)]
```





# Качество предсказания с новыми признаками

X Подсчитываем точность по классификации ПДЗ [0, 1-30, 31-90, 91-365]:

X Accuracy of train prediction is: **93.798%**

X Accuracy of test prediction is: **89.147%**



## Получаем значения для МаксПДЗ

### X Регрессия для МаксПДЗ:

X RMSE of train prediction is: **12.5 days of PDZ**

X RMSE of test prediction is: **138 days of PDZ**

### X Подсчитываем точность по классификации МаксПДЗ [0, 1-30, 31-90, 91-365]:

X Accuracy of train prediction is: **88.889%**

X Accuracy of test prediction is: **85.271%**



# Кластеризация компаний

- ✗ Интерес для будущего исследования представляет кластеризация компаний на несколько кластеров
- ✗ Есть видение, что при кластеризации компаний мы сможем лучше предсказывать просрочку, обучаясь внутри кластеров
- ✗ Для кластеризации предполагается использовать алгоритм DBScan





Евгений Васильев

[eugene.unn@gmail.com](mailto:eugene.unn@gmail.com)

[t.me/vasiliev\\_e](https://t.me/vasiliev_e)



Игорь Ситник

[goodday23@yandex.ru](mailto:goodday23@yandex.ru)

[t.me/sigal777](https://t.me/sigal777)

# THANKS!

Any questions?

You can find me at:

X [t.me/vasiliev\\_e](https://t.me/vasiliev_e)

X [eugene.unn@gmail.com](mailto:eugene.unn@gmail.com)

