ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Машинное обучение в анализе данных сельскохозяйственных предприятий

Харитонова Анна Евгеньевна

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

Имеются данные бухгалтерской отчетности по 20000 сельскохозяйственным предприятиям .
После объединения – 2000 строк, 138 столбцов.

[6]:
```
data
```

[6]:

Регион регистрации	Вид деятельности/ отрасль	ИДО	ИФР	ИПД	...	2021, Рентабельность затрат, %	2021, Доля себестоимости как процент от выручки, %	2021, Рентабельность прибыли до налогообложения и процентов - (EBITM), %	Рентаб актив
енинградская область	Выращивание прочих плодовых и ягодных культур	52	90.0	NaN	...	0.0520	0.9505	NaN	
Московская область	Разведение молочного крупного рогатого скота, ...	52	97.0	NaN	...	-0.1256	1.1436	-0.0394	
Москва	Рыбоводство пресноводное индустриальное	19	15.0	NaN	...	0.0582	0.9450	NaN	
Орловская область	Выращивание зерновых (кроме риса), зернобобовы...	18	11.0	NaN	...	0.3487	0.7414	NaN	
Тверская область	Лесоводство и прочая лесохозяйственная	18	90.0	NaN	...	-0.2330	1.3037	NaN	

[7]:
```
data.describe()
```

[7]:

Возраст компании, лет	ИДО	ИФР	ИПД	Уставный капитал, RUB	2021, Доходы, RUB	2021, Расходы, RUB	2021, Налоги, RUB
20000.000000	20000.000000	19997.000000	956.000000	1.999200e+04	1.997500e+04	1.997500e+04	1.961400e+04
14.204875	21.012100	41.530830	89.864017	2.405039e+07	2.275470e+08	1.874843e+08	1.102767e+07
8.359824	23.085069	32.362586	16.115042	3.069225e+08	2.435618e+09	2.350989e+09	5.407321e+07
0.500000	1.000000	3.000000	0.000000	8.000000e+00	-4.010000e+05	0.000000e+00	0.000000e+00
7.000000	2.000000	13.000000	85.000000	1.000000e+04	3.468000e+06	3.430000e+06	2.089560e+05
13.500000	11.000000	28.000000	97.000000	1.100000e+04	2.131900e+07	1.890900e+07	8.616635e+05
19.500000	34.000000	76.000000	100.000000	1.951195e+05	1.007485e+08	8.330800e+07	4.223042e+06
95.500000	99.000000	99.000000	100.000000	2.817760e+10	3.118717e+11	3.080237e+11	1.954704e+09

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Обработка пропусков

```
ИФР . Тип данных float64. Количество пустых значений 3, 0.01%.
ИПД . Тип данных float64. Количество пустых значений 19044, 95.22%.
Уставный капитал, RUB . Тип данных float64. Количество пустых значений 8, 0.04%.
2021, Доходы, RUB . Тип данных float64. Количество пустых значений 25, 0.12%.
2021, Расходы, RUB . Тип данных float64. Количество пустых значений 25, 0.12%.
2021, Налоги, RUB . Тип данных float64. Количество пустых значений 386, 1.93%.
2021, Период погашения кредиторской задолженности, дни . Тип данных float64. Количество пустых зна
чений 1923, 9.62%.
2021, Оборачиваемость кредиторской задолженности, разы . Тип данных float64. Количество пустых зна
чений 1808, 9.04%.
2021, Период оборота запасов, дни . Тип данных float64. Количество пустых значений 2873, 14.37%.
2021, Оборачиваемость запасов, разы . Тип данных float64. Количество пустых значений 3116, 15.58%.
2021, Период погашения дебиторской задолженности, дни . Тип данных float64. Количество пустых знач
ений 2024, 10.12%.
2021, Оборачиваемость дебиторской задолженности, разы . Тип данных float64. Количество пустых знач
ений 2351, 11.76%.
2021, Оборачиваемость основных средств, разы . Тип данных float64. Количество пустых значений 475
9, 23.79%.
2021, Период оборота основных средств, дни . Тип данных float64. Количество пустых значений 4098,
20.49%.
2021, Период оборота активов, дни . Тип данных float64. Количество пустых значений 65, 0.33%.
2021, Соотношение валовой прибыли к активам компании, % . Тип данных float64. Количество пустых зн
ачений 7068, 35.34%.
2021, Соотношение дебиторской задолженности к активам компании, % . Тип данных float64. Количество
пустых значений 2297, 11.48%.
2021, Доля рабочего капитала в активах компании, % . Тип данных float64. Количество пустых значени
й 162, 0.81%.
2021, Коэффициент оборачиваемости совокупных активов, % . Тип данных float64. Количество пустых зн
ачений 53, 0.27%.
2021, Коэффициент соотношения заемных и собственных средств, % . Тип данных float64. Количество пу
с]
```

```
[18]: # Замена переменных по этим столбцам не корректна (может сильно исказить результаты.
      # Удаление строк, содержащих пустые значения
      data_2 = data_1.dropna(axis=0, how='any')
      (data_2.shape, data_1.shape)

[18]: ((17821, 111), (20000, 111))
```

Расчет системы относительных показателей

```
# Выберем список относительных показателей, которые нам понадобятся для построения модели для прове
x_col_list2 = ['Прибыль в расчете на 1 работника', 'Возраст компании, лет','Оборотный капитал на 1
data_all=data_2[x_col_list2]
data_all.head()
```

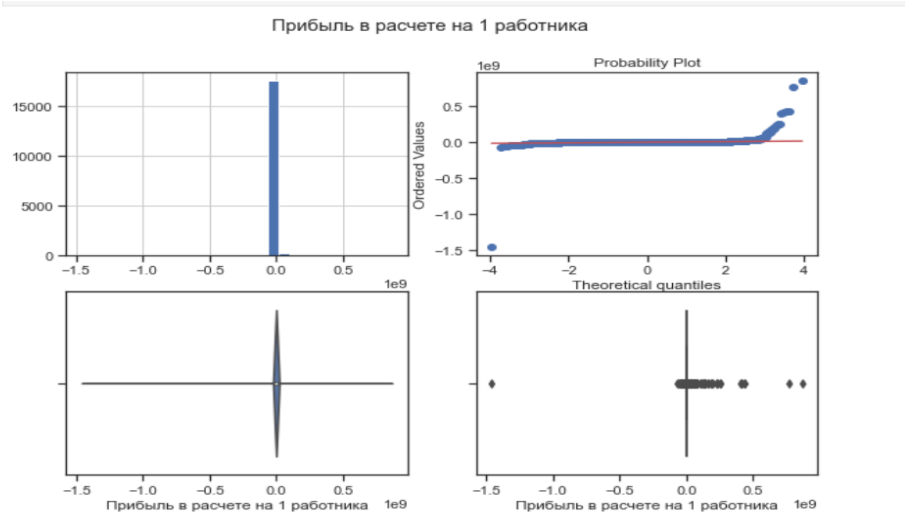
	Прибыль в расчете на 1 работника	Возраст компании, лет	Оборотный капитал на 1 работника	Чистые активы на 1 работника	Совокупный долг на 1 работника	Налоговая нагрузка	Выручка на 1 рубль активов	Налоги на 1 работника	Кре задо на 1
0	7.700000e+04	4.5	4.585000e+06	1.000000e+04	7.091000e+06	0.174699	0.219265	272007.000000	2.5
1	-2.612500e+05	2.5	2.667500e+05	7.007500e+05	2.137000e+06	0.040449	0.641001	73576.000000	4.1
2	5.550000e+04	20.5	4.197500e+05	4.197500e+05	2.202500e+05	0.096504	1.576953	97397.000000	2.1
3	1.456538e+06	7.5	1.119769e+06	2.005769e+06	3.150077e+06	0.027554	1.092621	155220.846154	2.1
4	-3.548000e+05	13.0	3.100000e+04	1.000000e+03	3.350000e+05	0.073065	3.476488	85346.700000	3.1

Прибыль в расчете на 1 работника
Возраст компании, лет
Оборотный капитал на 1 работника
Чистые активы на 1 работника
Совокупный долг на 1 работника
Налоговая нагрузка
Выручка на 1 рубль активов
Налоги на 1 работника
Кредиторская задолженность на 1 работника
Налоги на 1 руб. себестоимости
Фондовооруженность
Доля внеоборотных активов в общих
Доля рабочего капитала в активах компании, %
Коэффициент оборачиваемости совокупных активов, %
Соотношение чистого долга к капиталу, %
Коэффициент концентрации собственного капитала (автономии), %
Коэффициент маневренности собственных средств, %
Коэффициент обеспеченности собственными оборотными средствами, %
Рентабельность затрат, %
Доля себестоимости как процент от выручки, %
Рентабельность активов (ROA), %
Рентабельность капитала (ROE), %

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Обработка выбросов

Количество выбросов в столбце Прибыль в расчете на 1 работника : 2455
Количество выбросов в столбце Возраст компании, лет : 14
Количество выбросов в столбце Оборотный капитал на 1 работника : 2042
Количество выбросов в столбце Чистые активы на 1 работника : 1924
Количество выбросов в столбце Совокупный долг на 1 работника : 2031
Количество выбросов в столбце Налоговая нагрузка : 1957
Количество выбросов в столбце Выручка на 1 рубль активов : 2116
Количество выбросов в столбце Налоги на 1 работника : 2033
Количество выбросов в столбце Кредиторская задолженность на 1 работника : 2365
Количество выбросов в столбце Налоги на 1 руб себестоимости : 1681
Количество выбросов в столбце Фондовооруженность : 1577
Количество выбросов в столбце Доля внеоборотных активов в общих : 1
Количество выбросов в столбце 2021, Доля рабочего капитала в активах компании, % : 9
И др.

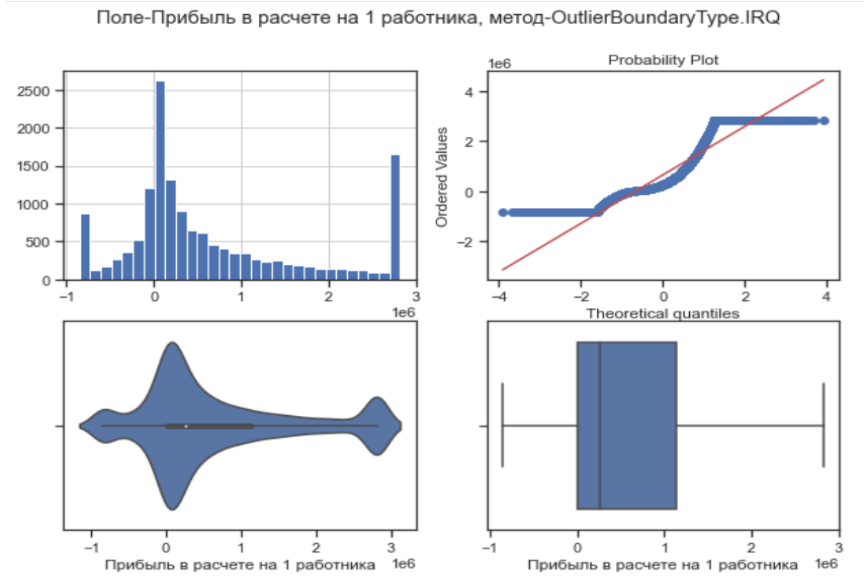


[24]:
```
# Основные статистические характеристики набора данных
data_all.describe()
```

[24]:

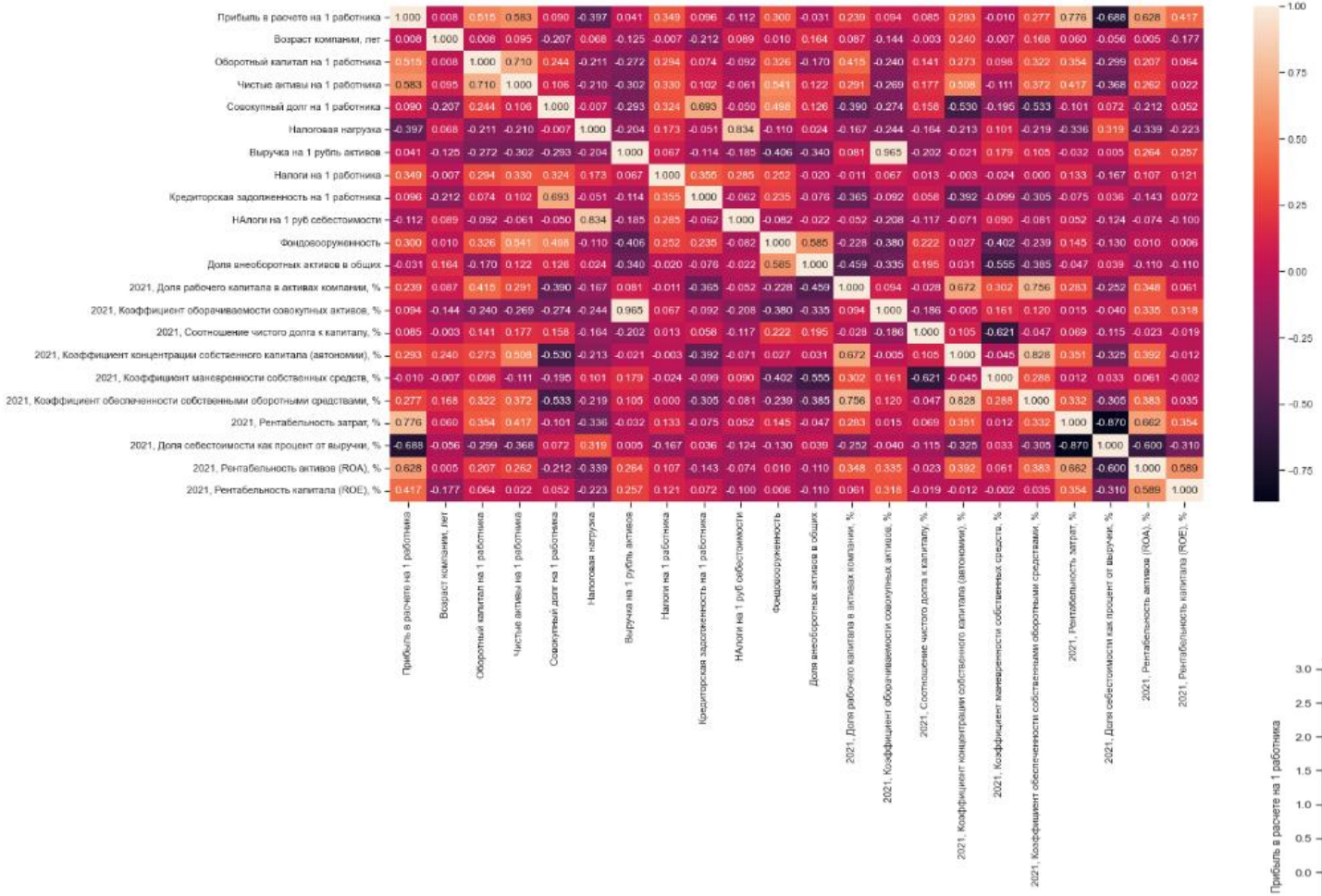
	Прибыль в расчете на 1 работника	Возраст компании, лет	Оборотный капитал на 1 работника	Чистые активы на 1 работника	Совокупный долг на 1 работника	Налоговая нагрузка	Выручка на 1 рубль активов	Нало раб
count	1.782100e+04	17821.000000	1.782100e+04	1.782100e+04	1.782100e+04	17821.000000	17821.000000	1.7821
mean	1.248437e+06	14.407862	1.959655e+06	3.177732e+06	1.020785e+07	0.354084	3.723376	3.3699
std	1.663880e+07	8.368516	8.968357e+07	7.846588e+07	1.174726e+08	6.175596	77.862137	2.2939
min	-1.456467e+09	1.000000	-8.137725e+09	-5.549908e+09	-1.352300e+07	0.000000	-7.164179	0.0000
25%	3.162791e+03	7.500000	2.462222e+05	2.370000e+05	3.540000e+05	0.026647	0.337654	7.0618
50%	2.409545e+05	14.000000	1.327208e+06	1.832000e+06	1.592545e+06	0.052810	0.614693	1.1740
75%	1.147667e+06	20.000000	3.552000e+06	5.197870e+06	5.135643e+06	0.101403	1.250256	2.2001
max	8.655915e+08	95.500000	9.953140e+08	1.326916e+09	1.001313e+10	569.739851	6231.000000	1.5601

8 rows × 22 columns

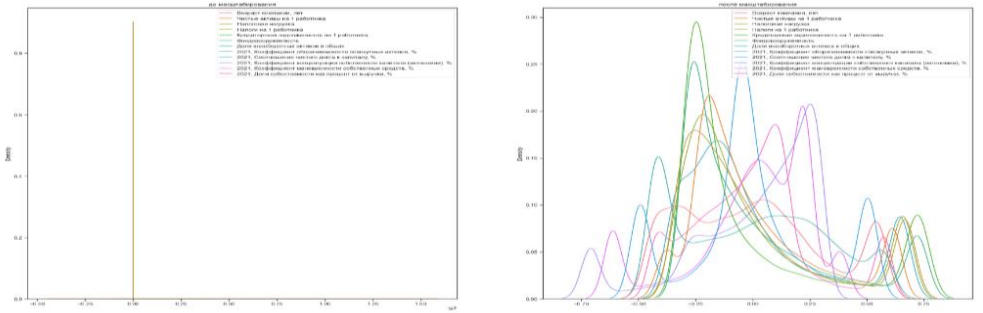


ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

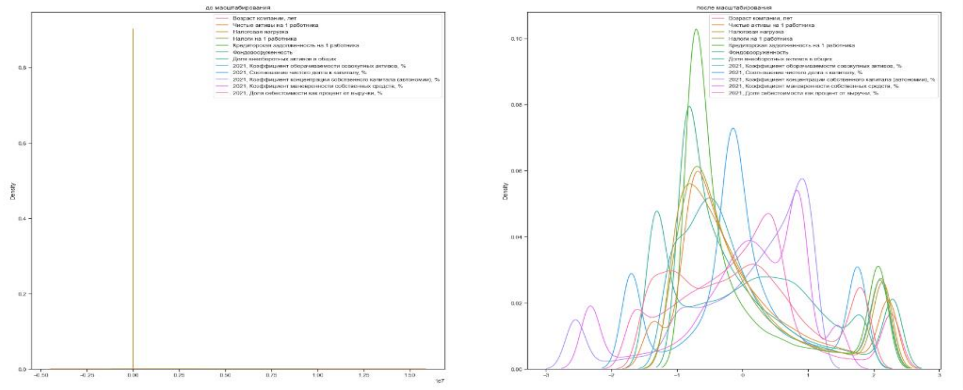
Подбор факторов

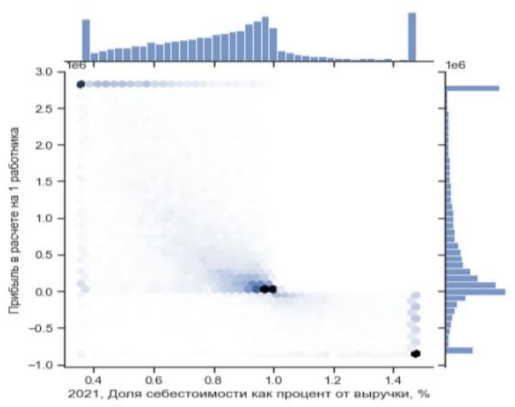


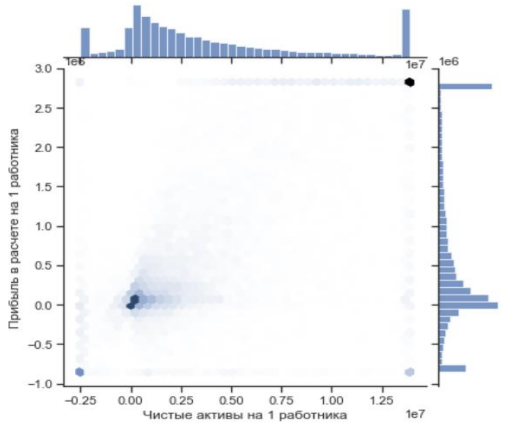
MinMaxScaler



StandardScaler





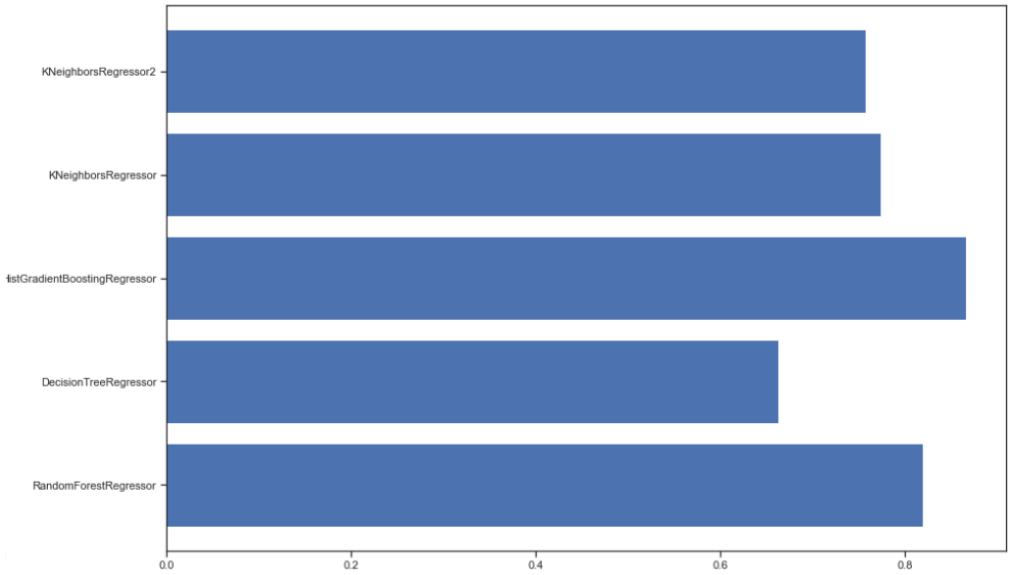


ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

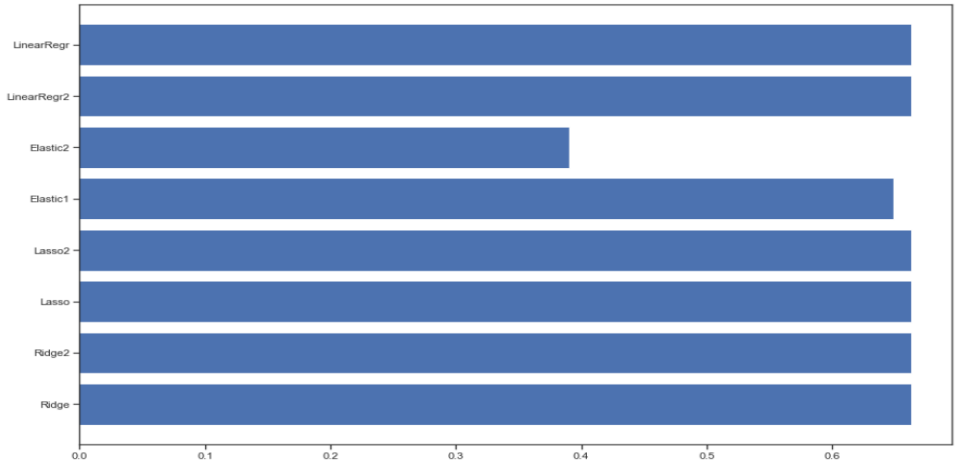
Разработка и обучение моделей

Значения признаков находятся в разных диапазонах, поэтому сделаем нормализацию данных (MinMaxScaler). Также проведем стандартизацию как дополнительный способ предобработки данных (StandardScaler)

	Модель регрессии	MAE	R2	MAPE
0	RandomForestRegressor	284793.8537812135	0.8191112290948995	2.4040733199565323
1	DecisionTreeRegressor	401504.77522422123	0.6631287782227941	3.3690619049305446
2	HistGradientBoostingRegressor	229917.24865967268	0.8657640283125339	2.9337421155684766
3	KNeighborsRegressor	293499.76760809514	0.773839313763896	2.7488765325224436
4	KNeighborsRegressor2	305743.8430911457	0.7570737457953616	2.6752949453851333



	Модель регрессии	MAE	R2	MAPE
0	Ridge	423749.251196	0.662625	7.602663
1	Ridge2	423762.812707	0.662615	7.599451
2	Lasso	423755.216277	0.662624	7.603404
3	Lasso2	423755.216810	0.662624	7.603404
4	Elastic1	446066.925240	0.648926	6.657384
5	Elastic2	616376.160531	0.390617	9.304927
6	LinearRegr2	423755.216089	0.662624	7.603404
7	LinearRegr	423755.216089	0.662624	7.603404



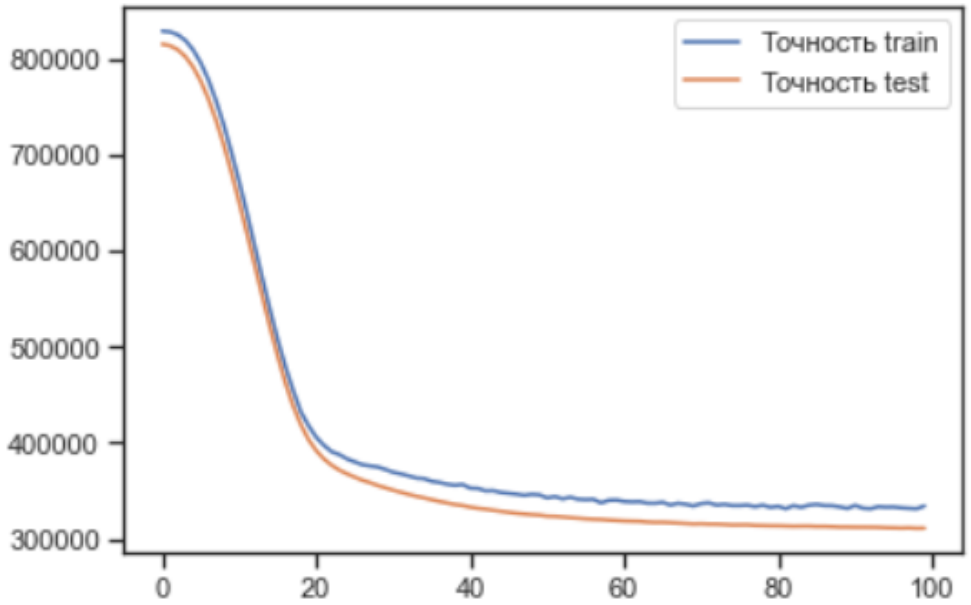
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

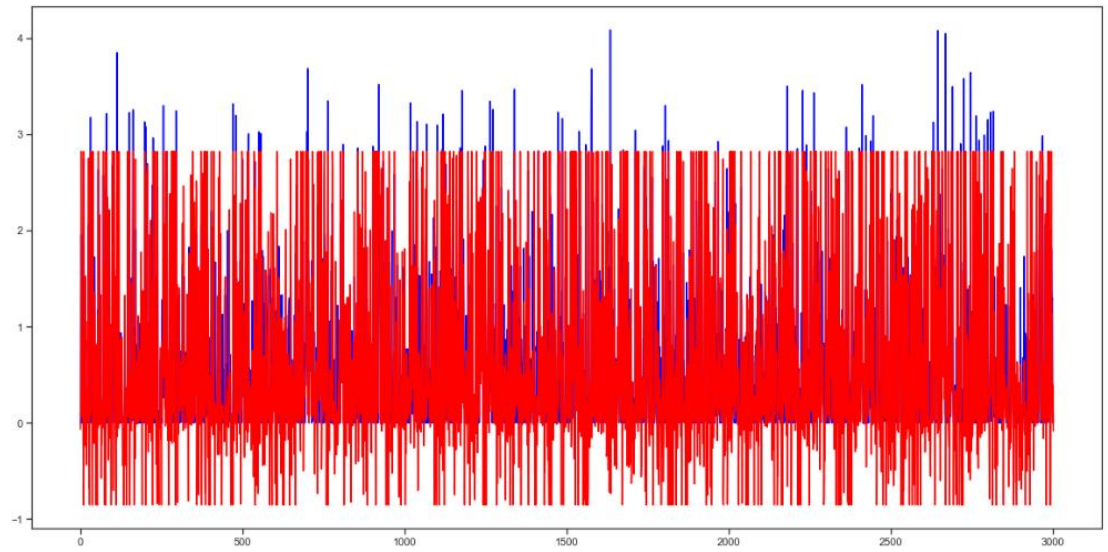
Нейронная сеть

[7]:
```
#создание метрики R2
from keras import backend as K

def coeff_determination(y_true, y_pred):
    SS_res =  K.sum(K.square( y_true-y_pred ))
    SS_tot = K.sum(K.square( y_true - K.mean(y_true) ) )
    return ( 1 - SS_res/(SS_tot + K.epsilon()) )
```

[61]:
```
# модель полносвязной нейронной сети для целевого параметра y1 relu выпрямленная линейная единица.
#ReLU математически определяется как F (x) = max (0, x). Другими словами, на выходе будет x, если x
modelNN_y1 = Sequential()
modelNN_y1.add(Dense(64, input_dim=12, activation = 'relu'))
modelNN_y1.add(Dropout(0.3))
modelNN_y1.add(Dense(32, activation = 'relu'))
modelNN_y1.add(Dense(1, activation = 'relu'))
#modelNN_y1.compile(optimizer='rmsprop', loss='mse', metrics=[coeff_determination])
modelNN_y1.compile(optimizer='adam', loss='mse', metrics=['mae', coeff_determination])
modelNN_y1.summary()
history = modelNN_y1.fit(Data_scaled_train_1,Data_scaled_train_y,
                    epochs=100,
                    validation_split=0.2,
                    verbose=2)
plt.plot(history.history['mae'], label = 'Точность train')
plt.plot(history.history['val_mae'], label = 'Точность test')
plt.xlabel = ('Epochs')
plt.ylabel = ('mae')
plt.legend()
plt.show()
```

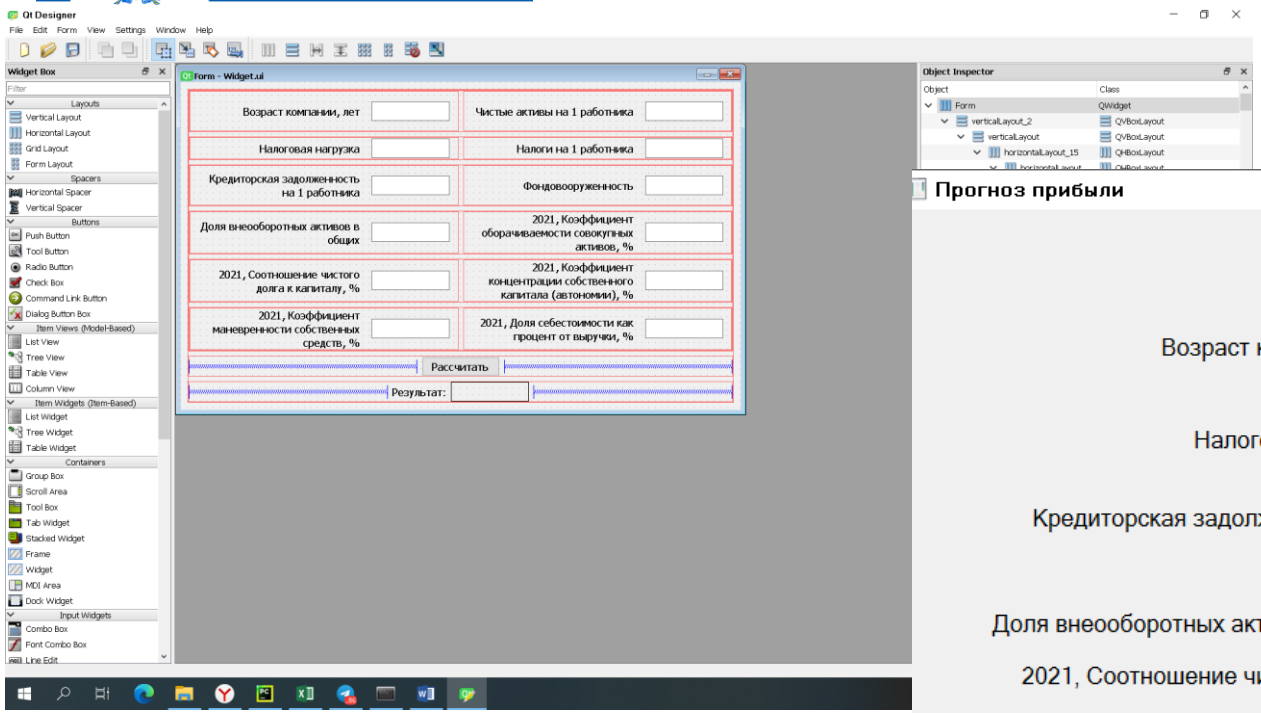


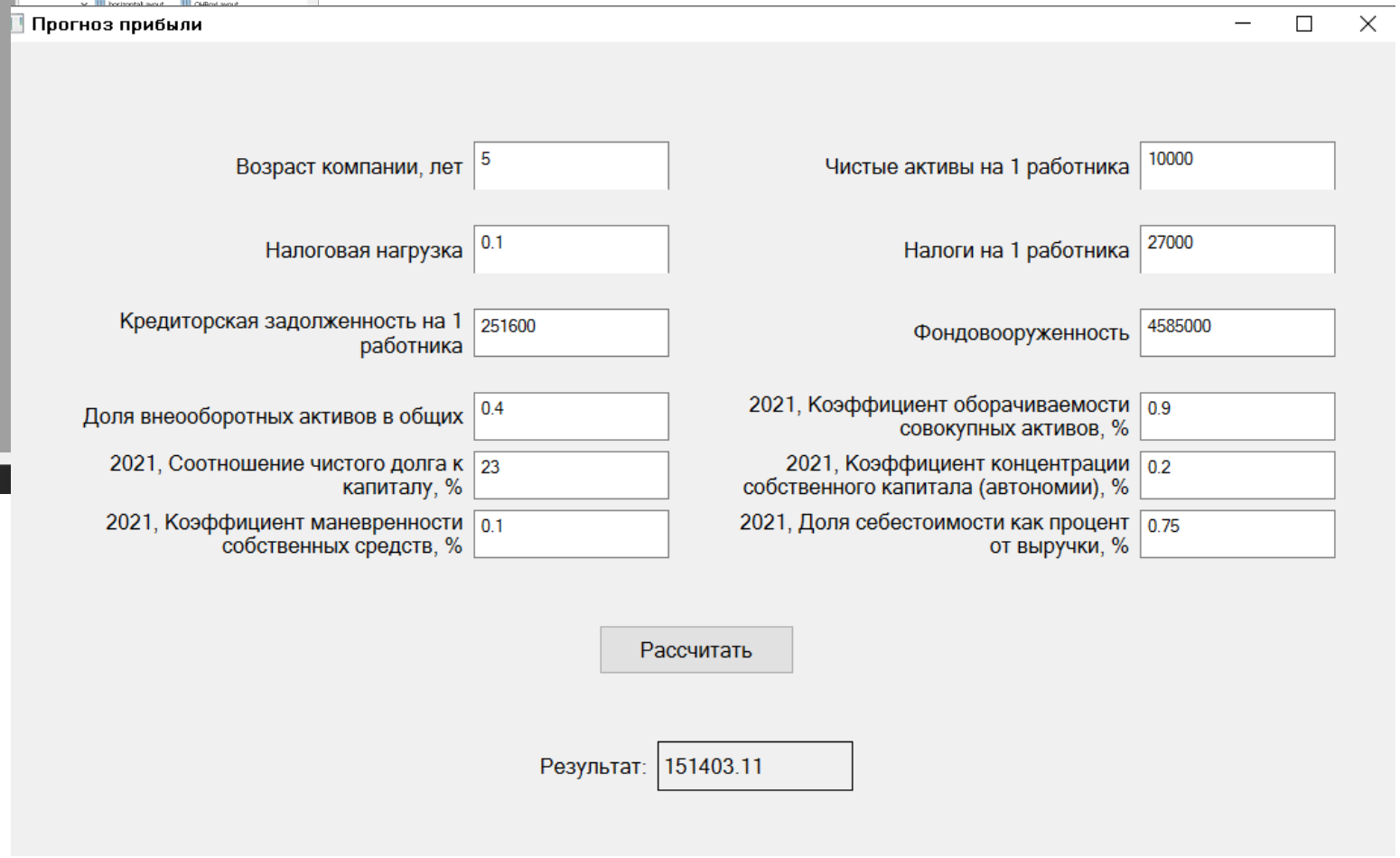


R2=0,77

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка приложения Qt





ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru