**Выполнил:**

студент 3 курса группы ПМ20-1

Джаводов Санджар

**Построение логит модели**

В качестве объекта исследования выступили сельскохозяйственные организации из Нижегородской области за период 2017-2020 гг. Для построения скоринговой модели воспользовался работой А.Н. Бобрышева и Р.В. Дебелого, где они выделили необходимые показатели для оценки финансового состояния предприятия.

*Предобработка данных:*

Начнем работу с построения новой таблицы где будут включены необходимые показатели по Нижегородской области. Были отобраны следующие показатели:

* Статус
* Регион регистрации
* 2017, Рентабельность капитала (ROE), %
* 2018, Рентабельность капитала (ROE), %
* 2019, Рентабельность капитала (ROE), %
* 2020, Рентабельность капитала (ROE), %
* 2017, Рентабельность продаж, %
* 2018, Рентабельность продаж, %
* 2019, Рентабельность продаж, %
* 2020, Рентабельность продаж, %
* 2017, Рентабельность активов (ROA), %
* 2018, Рентабельность активов (ROA), %
* 2019, Рентабельность активов (ROA), %
* 2020, Рентабельность активов (ROA), %
* 2017, Коэффициент обеспеченности собственными оборотными средствами, %
* 2018, Коэффициент обеспеченности собственными оборотными средствами, %
* 2019, Коэффициент обеспеченности собственными оборотными средствами, %
* 2020, Коэффициент обеспеченности собственными оборотными средствами, %
* 2017, Коэффициент текущей ликвидности, %
* 2018, Коэффициент текущей ликвидности, %
* 2019, Коэффициент текущей ликвидности, %
* 2020, Коэффициент текущей ликвидности, %
* 2017, Коэффициент быстрой ликвидности, %
* 2018, Коэффициент быстрой ликвидности, %
* 2019, Коэффициент быстрой ликвидности, %
* 2020, Коэффициент быстрой ликвидности, %
* 2017, Коэффициент абсолютной ликвидности, %
* 2018, Коэффициент абсолютной ликвидности, %
* 2019, Коэффициент абсолютной ликвидности, %
* 2020, Коэффициент абсолютной ликвидности, %
* 2017, Коэффициент оборачиваемости совокупных активов, %
* 2018, Коэффициент оборачиваемости совокупных активов, %
* 2019, Коэффициент оборачиваемости совокупных активов, %
* 2020, Коэффициент оборачиваемости совокупных активов, %
* Отрасль
* Округ

Далее преобразовал категориальные данные, столбцы “Отрасль” и “Округ”. В нашем датасете довольно много пропущеннх значений, что также необходимо обработать. В данной работе все пропущенные значения заменены на среднее значение, воспользовавшись средствами Python. Классы в нашом датасете несбалансированные, что в дальнейшем помешает воспользоваться метриками оценки логистической модели.

Для балансировки данных был применен алгоритм SMOTE (метод синтетической передискретизации меньшинства) в Python. Для того, чтобы устранить несколько признаков, которые сильно коррелировали между собой был применен алгоритм рекурсивного устранения признаков. Суть алгоритма заключается в многократном построении модели и выбора либо наилучшей, либо наихудшей функции, откладывания функции в сторону, а затем повторения процесса с остальными функциями. Этот процесс применяется до тех пор, пока не будут исчерпаны все объекты в наборе данных. Цель RFE состоит в том, чтобы выбирать объекты путем рекурсивного рассмотрения все меньших и меньших наборов объектов.

После данного алгоритма были удалены следующие показатели:

* 2017, Рентабельность продаж, %
* 2018, Рентабельность продаж, %
* 2019, Рентабельность продаж, %
* 2020, Рентабельность продаж, %
* 2017, Коэффициент быстрой ликвидности, %,
* 2018, Коэффициент быстрой ликвидности, %
* 2019, Коэффициент быстрой ликвидности, %
* 2020, Коэффициент быстрой ликвидности, %
* Округ

Теперь перейдем к построению моделей.

*Построение моделей:* Разделим весь наш датасет по годам и получим 4 новых датасета.

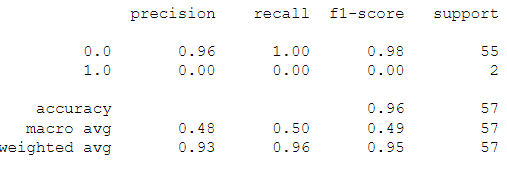
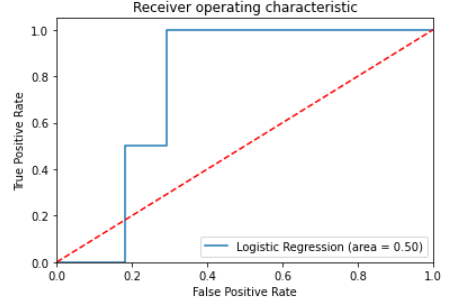
Для построения модели необходимо поделить выборку на обучающую и тестовую. В данной работе датасет поделен на обучающую и тестовую в соотношении 70 на 30.

***Регрессионная таблица логистической модели за 2017 год.***

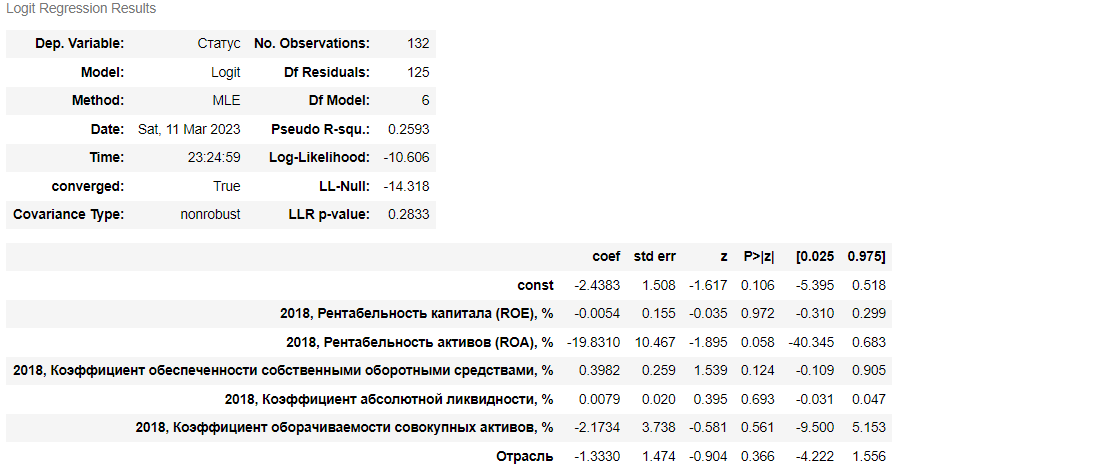


*Таблица 1*

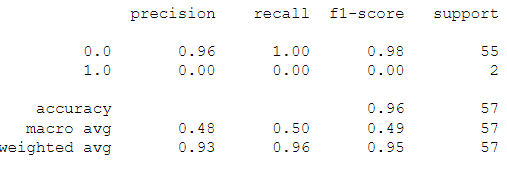
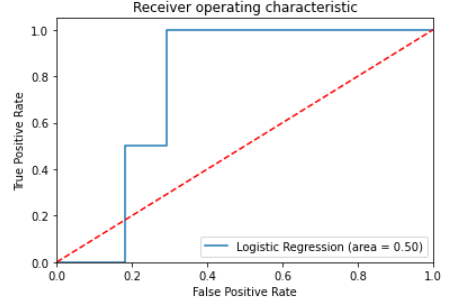
Для оценки модели задействованы такие метрики как: accuracy, precision, recall, f-1 score, support, confusion matrix:

1. Точность классификатора логистической регрессии на тестовом наборе: 0,96
2. Confusion matrix: [[55 0] [ 2 0]], результат говорит нам, что у нас есть 55 + 0 правильных предсказаний и 2 + 0 неправильных предсказания.
3. 
4. 

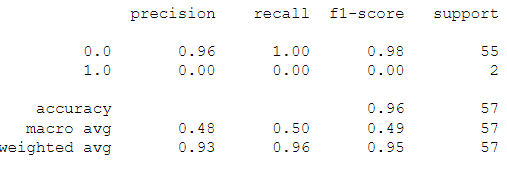
***Регрессионная таблица логистической модели за 2018 год.***



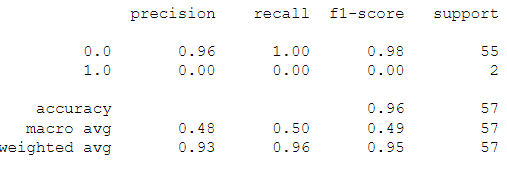
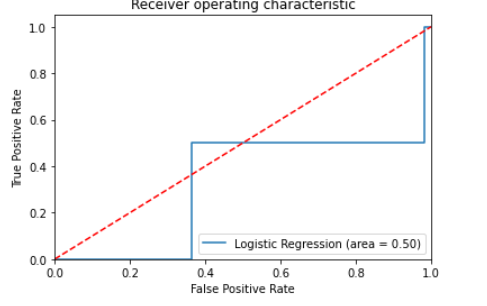
*Таблица 2*

1. Точность классификатора логистической регрессии на тестовом наборе: 0,96
2. Confusion matrix: [[55 0] [ 2 0]], результат говорит нам, что у нас есть 55 + 0 правильных предсказаний и 2 + 0 неправильных предсказания.
3. 
4. 

***Регрессионная таблица логистической модели за 2019 год.***

1. Точность классификатора логистической регрессии на тестовом наборе: 0,96
2. Confusion matrix: [[55 0] [ 2 0]], результат говорит нам, что у нас есть 55 + 0 правильных предсказаний и 2 + 0 неправильных предсказания.
3. 
4. 

***Регрессионная таблица логистической модели за 2020 год.***

1. Точность классификатора логистической регрессии на тестовом наборе: 0,96
2. Confusion matrix: [[55 0] [ 2 0]], результат говорит нам, что у нас есть 55 + 0 правильных предсказаний и 2 + 0 неправильных предсказания.
3. 
4. 

С большими допущениями можно считать, что чем больше показатель AUC, тем лучшей прогностической силой обладает модель. Однако следует знать, что:

* показатель AUC предназначен скорее для сравнительного анализа нескольких моделей;
* AUC не содержит никакой информации о чувствительности и специфичности модели.

По приведенной выше статистике, можно сделать вывод что логит-модель достаточно хорошая, хоть и последний показатель близок к неудовлетворительной отметке (показатель AUC), но выше приведена информация, которая обосновывает вывод. Недостаток модели заключался бы в доминировании одного класса над другим, но в разделе “предобработка данных” мы решили данную проблему.

# **Заключение.**

В данной работе были построены логит модели для Нижегородского края за период 2017-2020. Данные были отфильтрованы, пред обработаны и сбалансированы. Считаю, что для построения более устойчивой модели необходимо увеличить географию датасета, то есть включить остальные регионы.