Company Bankruptcy Prediction

Тардова Александра, Панов Максим

11 ноября 2023 г.

Аннотация

Мы анализировали данные о банкротстве компаний с ресурса kaggle: доступен по ссылке

1 Использованные библиотеки

Работа с данными:

- pandas
- numpy

Графики:

- plotly
- seaborn
- matplotlib.pyplot

Построение моделей машинного обучения и метрики качества:

- sklearn

2 Работа с данными

2.1 Выбор данных

В выданном датасете оказалось 96 колонок данных. Если использовать для обучения все, то можно получить эффект переобучения, ведь некоторые могут быть нерелевантны с точки зрения предсказания банкротства. Так, для построения модели логистической регрессии были использованы отношение денежнего потока к обязательствам, процентов долга к активам, текущих обязательств к активам и некоторые другие.

2.2 Использованные модели и доказательство ее релевантности

Мы считаем, что можно предсказывать банкротство конкретной компании по схожим с ней по характеристикам компаниям, поэтому мы решили использовать метод k-ближайших соседей, где схожесть вычисляется как расстояние между объектами на основе характеристик. Второй метод, которых мы использовали - это логистическая регрессия. Мы сочли этот метод подходящим, так как он отлично работает в задачах классификации, к которым относится и наша. То есть, наша вторая модель оценивает, является ли компания банкротом по нескольким заданным параметрам, исходя из данных, на которых она обучалась.

2.3 Описание графиков

Как было отмечено выше, датасет содержит слишком много параметров, поэтому мы использовали метод feature_importances_ из библиотеки sklearn. Далее взяли 10 самых важных и обучили модель на них. Они представлены на графике из plotly, с возможностью слайдинга по OX.

Для разнообразия, второй график было решено делать для оценки эффективности работы второй модели. Мы использовали метод ROC кривой (оценка кол-ва правильно вынесенных положительных оценок к неверно вынесенным положительным) как наиболее подходящий для нашей задачи. Второй график также был реализован с помощью plotly, еще было использовано несколько методов из sklearn.

3 Описание полученных результатов

Для метода k-ближайших соседей мы обучили модель на характеристиках, подобранных feature_importances_.

Нам удалось добиться хороших mse и mae (0.05 и 0.03 соответственно), указывающих на низкое отклонение между фактическими и предсказанными значениями. Тем не менее, метрика $R^2, 26\%$.

В методе логистической регрессии было использовано шесть показателей,про которые уже было упомянуто в Выборе данных. Модель продемонстрировала отличные результаты с точностью в 96%, для оценки также использовались матрица путаницы и ROC кривая, на которых модель тоже показала себя хорошо.