# 3 Выбор данных.

## 3.1 Набор данных.

В данной статье, данные были взяты с сайта UCI machine learning repository. Изначально они были извлечены с ресурса Emerging Markets Information Services (EMIS, securities.com), который представляет собой базу данных, содержащую информацию о развивающихся рынках по всему миру. Обанкротившиеся компании были проанализированы в период 2000-2012 годов, в то время как все еще действующие компании были оценены с 2007 по 2013 год Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Основные метаданные готового набора данных представлены в таблице 1.

Таблица 1. Основные характеристики набора данных.

|  |  |
| --- | --- |
| Название | Описание |
| Характеристика набора данных | Многофакторный |
| Характеристика показателей | Реальные |
| Тип задачи | Классификация |
| Количество объектов | 10503 |
| Количество показателей | 64 |
| Пропущенные значения | Да |
| Сфера деятельности | Бизнес/Экономика |

В каждом наборе данных содержатся 64 финансовых показателя + класс прогнозирования (0 – не банкрот, 1 – банкрот)

*1st year* – данные содержат финансовые показатели с 1-го года прогнозируемого периода и метку соответствующего класса, которая указывает на статус банкротства через 5 лет.

*2nd year* – данные содержат финансовые показатели за 2-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 4 года.

*3rd year* – данные содержат финансовые показатели за 3-ий год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 3 года.

*4th year* – данные содержат финансовые показатели за 4-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 2 года.

*5th year* – данные содержат финансовые показатели за 5-й год прогнозируемого периода и метку соответствующего класса, которая указывает на состояние банкротства через 1 год.

Расшифровка финансовых показателей финансовых показателей представлена в таблице 2.

Таблица 2 – Описание финансовых показателей

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Описание** | **ID** | **Описание** |
| X1 | Чистая прибыль/Суммарные активы | X33 | Операционные расходы/ Краткосрочные обязательства |
| X2 | Сумма всех обязательств/Суммарные активы | X34 | Операционные расходы/ Суммарные обязательства |
| X3 | Текущие активы/ Суммарные активы | X35 | Прибыль от продаж/ Суммарные активы |
| X4 | Текущие активы/ Краткосрочные обязательства | X36 | Суммарные продажи/ Суммарные активы |
| X5 | [(Наличные ДС + краткосрочные инвестиции + дебиторская задолженность - Краткосрочные обязательства)/ (операционные расходы - амортизация)] | X37 | (Текущие активы – запасы)/ Долгосрочные обязательства |
| X6 | Нераспределенная прибыль/ Суммарные активы | X38 | (Постоянный капитал)/ Суммарные активы |
| X7 | Прибыль до налогов и процентов/ Суммарные активы | X39 | Прибыль от продаж/ продажи |
| X8 | Балансовая стоимость капитала/ Суммарные обязательства | X40 | (Текущие активы – запасы – дебиторская задолженность)/ Краткосрочные обязательства |
| X9 | Продажи/ Суммарные активы | X41 | Суммарные обязательства/(( Прибыль от операционной деятельности + амортизация) \* (12/365)) |
| X10 | Капитал/ Суммарные активы | X42 | Прибыль от операционной деятельности/ Продажи |
| X11 | (Валовая прибыль + иные активы и транзакции + финансовые затраты)/ Суммарные активы | X43 | ротация дебиторской задолженности + товарооборот (в днях) |
| X12 | Валовая прибыль/ Краткосрочные обязательства | X44 | (дебиторская задолженность \* 365) / Продажи |
| X13 | (Валовая прибыль + амортизация) / Продажи | X45 | Чистая прибыль/ запасы |
| X14 | (Валовая прибыль + выплаты по кредиту)/ Суммарные активы | X46 | (Текущие активы - запасы) /Краткосрочные обязательства |
| X15 | (Суммарные обязательства \* 365)/ (Валовая прибыль + амортизация) | X47 | (Запасы \* 365)/ Издержки проданной продукции |
| X16 | (Валовая прибыль + амортизация)/ Суммарные обязательства | X48 | прибыль до вычета процентов, налогов и амортизации (Прибыль от операционной деятельности – амортизация) / Суммарные активы |
| X17 | Суммарные активы/ суммарные обязательства | X49 | прибыль до вычета процентов, налогов и амортизации (Прибыль от операционной деятельности – амортизация) / Продажи |
| X18 | Валовая прибыль/ Суммарные активы | X50 | Текущие активы/ Суммарные обязательства |
| X19 | Валовая прибыль/ Продажи | X51 | Краткосрочные обязательства/ Суммарные активы |
| X20 | (Инвентарь \* 365) / Продажи | X52 | (Краткосрочные обязательства\*365)/ Издержки на проданную продукцию |
| X21 | (продажи (n) / продажи (n-1)) | X53 | Капитал/ Основные средства |
| X22 | Прибыль от операционной деятельности/ Суммарные активы | X54 | Постоянный капитал/ основные средства |
| X23 | Чистая прибыль/ Продажи | X55 | Оборотный капитал |
| X24 | Валовая прибыль (за 3 года)/ | X56 | (Продажи – издержки проданной продукции) / Продажи |
| X25 | (Капитал – акционерский капитал)/ Суммарный капитал | X57 | (Текущие активы - запасы – краткосрочные обязательства)  / (Продажи – Валовая прибыль - амортизация) |
| X26 | (Чистая прибыль + амортизация)/ Суммарные обязательства | X58 | Суммарные издержки/ Суммарные продажи |
| X27 | (Прибыль от операционной деятельности)/ Финансовые расходы | X59 | Долгосрочные обязательства/ Капитал |
| X28 | Оборотный капитал/ основные средства | X60 | Продажи/ Запасы |
| X29 | Логарифм от суммарных активов | X61 | Продажи/ Дебиторская задолженность |
| X30 | (Суммарные обязательства – наличные ДС)/продажи | X62 | (Краткосрочные обязательства \* 365)/ продажи |
| X31 | (Валовая прибыль + выплаты по кредитам)/ продажи | X63 | Продажи/ Краткосрочные обязательства |
| X32 | (Текущие обязательства \* 365)/ затраты на проданную продукцию | X64 | Продажи/ Основной капитал |

## 3.2 Предварительная обработка данных

Используемый для исследования набор данных имеет 2 проблемы:

* Неполноценность данных (большое количество пропущенных значений);
* Несбалансированность классов.

Для решения первой проблемы использовалась техника замены пропущенных величин с помощью среднего арифметического значения.

Так как данные являются несбалансированными (успешных предприятий намного больше, чем предприятий банкротов), для исправления дисбаланса была использована техника Synthetic Minority Oversampling Technique (SMOTE). Класс меньшинства дополняется новыми объектами, путем взятия каждого примера выборки и введения искусственных примеров вдоль отрезков, соединяющих между любой или все k-ближайших соседей. В зависимости от того какое количество дополнения набора данных необходимо, объекты из k-ближайших соседей выбираются случайным образом Nitech V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer (2002).

# 4 Эксперимент.

## 4.1 Цель эксперимента

Целью данного эксперимента является определение лучшей по точности прогнозирования модели и оценка важности финансовых показателей. Мы приняли во внимание следующие традиционные модели и алгоритмы машинного обучения в задачах классификации:

* Logistic Regression (LR),
* Decision Tree Classifier (DT),
* Random Forest Classifier (RDF),
* Extreme gradient boosting classifier (XGboost),
* Natural gradient boosting classifier (Ngboost).

Для каждой модели были определены оптимальные гиперпараметры с помощью техник Grid Search и Random Grid Search. Также использовалась 10-ти блочная перекрестная проверка.

Для предварительной обработки данных использовались следующие библиотеки языка программирования python: imblearn, sklearn, numpy и pandas.При тестировании различных алгоритмов также были использованы такие библиотеки как: sklearn и xgboost.